

NEURAL CORRELATES OF ECONOMIC
AND MORAL DECISION-MAKING

Thesis by

Cédric Robert Anen

In Partial Fulfillment of the Requirements for the
degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2007

(Defended May 18, 2007)

© 2007

Cédric Robert Anen

All Rights Reserved

ACKNOWLEDGEMENTS

Having constantly worked in electrical engineering, my encounter with neuroeconomics was rather a coincidence. After having spent the first year of my Ph.D. taking classes in electrical engineering at Caltech, in summer 2002 I came across a newspaper article about the inconsistencies in choices when rewards are presented at different points in time, a phenomenon known as time discounting in behavioral economics. Researchers were trying to uncover the neural correlates of evaluating future rewards, and my interest was immediately sparked. How great would it be if we could look into people's brains and analyze and explain the greatest mysteries of human behavior?! Although neuroeconomics was at that time a brand new science, I was extremely happy to find out that several professors at Caltech were already working in that field, and I was able to join Steve Quartz's lab in fall 2002.

I have spent wonderful years at Caltech, both on a personal and intellectual level. In my eyes no other place provides a better research environment to students and faculty. The reason is twofold: first of all, the relatively small size of Caltech allows different departments to work closely together. A lot of faculty have positions in two or more departments, and the students in their research groups also have diversified backgrounds. I believe that the research sparked by those interdisciplinary collaborations is at the core of scientific success. Secondly, no other campus has a concentration of such talented professors and students. Everyone at Caltech is bright and alert, and it seems that no-one ever gets tired of having discussion about research.

Most of all I would like to thank my advisor and mentor Steve Quartz for his continuous assistance and intellectual stimulation throughout my years at Caltech. I enjoyed working together with him and will definitely miss our discussions and speculations about brain activations. I will always admire his talent to see connections between seemingly unrelated results and to make everything fit together. I would also like to thank the other professors that I have collaborated or interacted with, and in particular Peter Bossaerts, Colin

Camerer, Read Montague (Baylor College of Medicine), and John O'Doherty. I would also like to thank the electrical engineering professors on my thesis committee, Shuki Bruck, Robert McEliece, and PP Vaidyanathan, for supporting me in my choice to pursue research that is not so typical for an electrical engineer.

A big thank you goes out to all the wonderful people at Caltech that I have worked with, discussed scientific problems with or just had fun with, in particular Ulrik Beierholm, Meghana Bhatt, Tony Bruguier, Ilja Friedel, Alan Hampton, Vanessa Heckman, Ming Hsu, Shreesh Mysore and Kerstin Preuschoff. All of you have made my time at Caltech special and it is with great regret that I will have to leave you.

Apart from science, my other love is fencing, and the Caltech Fencing Team has served as an important balance in my life. It was great hanging out and fencing with all of you, and in particular my friends and teammates: George, Yann, Joe, Randy, Christine, Rebecca, Kathy, Ken, Haomiao, Laura, and Steve. Caltech also offered me an immense opportunity by being able to coach the Caltech fencing team in 2004-5 and 2005-6, and to be the head coach of the team in 2006-7. This allowed me to experience the joy of teaching and to share my love of fencing with other people.

Last, but not least, I would also like to thank my family for the tremendous emotional support and encouragement they have provided. Although it must not have been easy for my parents Robert and Rita to have their oldest son study 6,000 miles away and to see him only once or twice a year, they strongly supported me in my choice to do my Ph.D. at Caltech. I cannot thank them enough. I would also like to thank my brother Adrien and my sister Nadège for staying in touch with me on a daily basis—their e-mails were definitely a welcome distraction on hard days of work!

ABSTRACT

Our daily lives are shaped by a series of decision processes, ranging from very unimportant choices to life-changing judgments. The complexity of the decision processes increases tremendously when the decision-making takes place in a social context, i.e., when other human beings are directly involved in the decision. In such conditions the decision-maker not only tries to maximize his own utility, but also needs to take into account the interdependent nature of the situation. Information about others' preferences, characteristics, and actions play an important role, and need to be thoroughly evaluated and predicted before making a decision. In this thesis we explore the neural correlates of two different types of social decision-making.

In the first experiment I investigate economic decision-making in the context of a two-player social exchange game. In order to maximize their overall and personal earnings, players need to cooperate and build up a trust relationship with their partner. Synchronized neural data is recorded from the two interacting brains using functional magnetic resonance imaging. In this thesis I present four main findings: (i) the neural correlates of strategic uncertainty and how it can be used to predict a player's future strategic choice; (ii) the dynamic interaction of the brains of two interacting players; (iii) the neural correlates of trust and its development over the course of the game; and (iv) how the brain distinguishes between one's own actions and those of another person.

The second experiment investigates the neural basis of moral decision-making and other-regarding preferences. Subjects have to make a morally difficult decision between helping two groups of children while trading off between efficiency and equity. By parametrically varying these variables, I show how two brain structures, the insula and the caudate, are actively involved in the decision-making process.

Taken together the results presented in this thesis shed some light on how our brain evaluates social situations, and how it uses social measures such as trust, agency, strategic interaction, and fairness to make decisions.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Table of Contents.....	vi
List of Illustrations and Tables	viii
 Chapter I: Introduction	 1
I.1. Neuroeconomics.....	2
I.2. Experiments in Neuroeconomics.....	6
 Chapter II: Basics of fMRI	 9
II.1. MRI Physics.....	9
II.2. Measuring the BOLD Signal.....	13
II.3. fMRI Data Acquisition	15
II.4. fMRI Data Preprocessing	17
II.5. fMRI Data Analysis	21
 Chapter III: Neural Correlates of Economic Decision-Making	 27
III.1. Background.....	27
III.1.1. Behavioral Game Theory	27
III.1.2. The Trust Game.....	31
III.1.3. Previous Studies	34
III.2. Experimental Design and Methods.....	35
III.2.1. Task	35
III.2.2. Subjects.....	36
III.2.3. Experimental Setup	36
III.2.4. fMRI Data Acquisition and Preprocessing.....	37
III.3. Behavioral Results.....	39
III.4. Strategic Uncertainty and Prediction Analysis.....	42
III.4.1. Background	42
III.4.2. GLM Analysis: New vs. Known Information.....	45
III.4.3. ROI Analysis: Correlation between Later Decisions and Hrf	47
III.4.4. Signal Magnitude Encodes Strategic Uncertainty.....	50
III.4.5. Signal Magnitude Predicts Future Strategic Choice	51
III.4.6. Discriminant Analysis of Predictive Accuracy	53
III.4.7. Discussion and Conclusion	54
III.4.8. Methods	58
III.5. Dynamic Cross-Brain Analysis.....	62
III.5.1. Background	62
III.5.2. Dependency Measures	63
III.5.3. Methods	69

III.5.4. Results	70
III.5.5. Discussion and Conclusion	76
III.6. Trust & Reciprocity	80
III.6.1. Background	80
III.6.2. Reciprocity Predicts Trust.....	81
III.6.3. “Intention to Trust” Signals	83
III.6.4. Model Building of Partner: Cross-Brain Analysis	83
III.6.5. Discussion and Conclusion	87
III.6.6. Methods	90
III.7. Agency Attribution.....	91
III.7.1. Background	91
III.7.2. Cross-Cingulate PCA Analysis.....	93
III.7.3. Differential “Own” and “Other” Responses.....	94
III.7.4. Agent-Specific Responses Disappear in Control Experiments.....	97
III.7.5. Cingulate Pattern Remains Constant for Several Variables	98
III.7.6. Discussion and Conclusion	100
III.8. Conclusions	103
Chapter IV: Neural Correlates of Moral Decision-Making.....	104
IV.1. Background	104
IV.2. Experimental Design and Methods	110
IV.2.1. Task	110
IV.2.2. Subjects	111
IV.2.3. Experimental Setup.....	111
IV.2.4. fMRI Data Acquisition and Analysis	114
IV.3. Behavioral Measures and Behavioral Results.....	115
IV.3.1. The Gini Coefficient	115
IV.3.2. Measures of Efficiency, Equity, and Utility.....	117
IV.3.3. Behavioral Data	118
IV.3.4. Act/Omit Differences	120
IV.3.5. Inequity Aversion Models	121
IV.4. fMRI Results	124
IV.4.1. Difference Between Allocations.....	124
IV.4.2. Correlates of Efficiency, Equity, and Utility in Take Trials..	126
IV.4.3. Correlates of Efficiency, Equity, and Utility in Give Trials..	127
IV.5. Discussion and Conclusions	129
IV.5.1. Insula Activations	130
IV.5.2. Caudate Activations	132
IV.5.3. Conclusions	133
IV.6. Methods	134
Chapter V: Conclusions	136
Bibliography	138

LIST OF ILLUSTRATIONS AND TABLES

<i>Figures</i>	<i>Page</i>
1. Available Choices in the Ellsberg Paradox	3
2. Spatiotemporal Resolution of Brain Recording Techniques	7
3. Hydrogen Nuclei in a Magnetic Field	10
4. Effect of A RF Pulse on the Net Magnetization.....	11
5. T1-Weighted High-Resolution Anatomical MR Image.....	13
6. T_2^* -Weighted Functional MR Image.....	14
7. Shape of the Hemodynamic Response Function.....	15
8. Full-Body Human 3-T MRI Scanner at Caltech	16
9. Normalization Procedure	19
10. Spatial Smoothing Procedure.....	20
11. Sample Time-Series of a Voxel Within the Brain.....	21
12. Effect of the Convolution with an Hrf.....	23
13. fMRI Activations in Glass-Brain and on SPM Map	24
14. The Ultimatum Game.....	27
15. The Prisoner's Dilemma	28
16. The Public Goods Game	29
17. The Trust Game.....	30
18. Fairness in the Trust Game	32
19. Hardware Setup of the Trust Game	35
20. Timeline of 1 Round of the Trust Game	37
21. Example #1 of Monetary Exchange	38
22. Example #2 of Monetary Exchange	39
23. Example #3 of Monetary Exchange	39
24. Average Investment and Repayment Ratios	40
25. Activations in the Trustee Brain	45
26. Split-up of the Behavioral Space	46

27. Time-Courses Predict Strategic Choice.....	48
28. Analysis of Signal Magnitudes	51
29. Performance of the Prediction Analysis	53
30. Activations in the Investor Brain	58
31. Distribution of the Mutual Information	67
32. Activations in the Investor Brain	70
33. Activations in the Trustee Brain	70
34. Game Dynamics	72
35. Cross-Round Dynamics in BA9	73
36. Cross-Round Dynamics in Primary Visual Cortex	76
37. Correlates of Reciprocity in a Multi-Round Economic Exchange	81
38. Correlograms of the “Intention to Trust”	84
39. Neural Correlates of Reputation Building in the Trustee Brain	85
40. Model Building in the Trustee Brain	86
41. Cingulate Segmentation	93
42. Cross-Cingulate Correlations	94
43. Agent-Specific Responses.....	95
44. Cingulate Pattern of “Me” and “Not Me”	98
45. Classical Dilemmas in Moral Decision-Making	105
46. Activations in Personal Moral Dilemmas	107
47. Example of a Child’s Biography	111
48. Timeline of the Moral Decision-Making Task.....	112
49. Graphical Representation of the Gini Coefficient.....	114
50. Wealth Distribution in the World	115
51. Subject Behavior in the Moral Task	118
52. Act/Omit Differences in the Moral Task.....	120
53. Coefficients in the Inequity Aversion Model	121
54. Repartition of α_{Take} Across Age	122
55. Difference Between Allocations	123
56. Time-Courses in Bilateral Insula in Take Trials	124

57. Delta Equity in Take Trials	125
58. Delta Utility in Take Trials	126
59. Delta Equity in Give Trials	127
60. Delta Efficiency in Give Trials	128
61. Interpretation of Delta Equity	130

<i>Tables</i>	<i>Page</i>
1. Investment > Repayment Regions for the Trustee	44
2. Repayment > Investment Regions for the Investor	58
3. Summary of Activations at the Revelation Screens	71
4. Summary of Game Dynamics	74
5. Allocation of Meals	110
6. Activations in the Insula for $ \Delta M $	123
7. Activations in the Insula for ΔG During Take Trials	125
8. Activations in the Caudate for ΔG During Give Trials	127

Chapter 1

INTRODUCTION

All living organisms face situations that require evaluating various alternatives and making decisions that are often critical to their survival. For example, a hunting tiger in the grasslands needs to decide whether it is sufficiently close to its prey to attack, or whether it should try to sneak closer at the risk of being seen. Such choices are typically associated with factors such as risk, uncertainty, reward, or punishment, which need to be estimated in order to make the best possible decision. As most organisms learn through experience, their decision-making processes are continuously updated and optimized, and eventually lead (after enough exposure or training) to a set of rules and actions that define the organism's behavior.

Humans also make such decisions, may they be relatively trivial such as picking between two different drinks in a bar, or more meaningful such as deciding whether or not to accept a job offer. What distinguishes us from most other organisms though is that some of our decisions are made with respect to outcomes that satisfy more than just the basic “animal” needs (e.g., hunger, thirst, sleep, reproduction). More specifically, many of our decisions have some economic value (e.g., money, time, power) or moral value (e.g., love, trust, respect) attached to them. Yet independent of the nature of that value, the common point underlying all decision-making processes is the need to design a strategy that allows us to make the best possible choice. This thesis discusses some aspects of the neural basis of economic and moral decision-making in the human brain.

I.1. Neuroeconomics

Human decision-making and choice theory have been thoroughly investigated in a variety of fields. Cognitive psychologists attempt to describe and understand behavior and mental processes by recording behavioral variables and psychometric measurements in controlled laboratory experiments. Behavioral economists aim to understand how human and social cognitive and emotional biases affect economic decisions, and they do this by designing a whole set of experiments and creating models to predict human choice. Cognitive neuroscientists are concerned with the neural substrates of mental processes and their behavioral manifestations. Traditionally, collaborations between these closely related fields have been relatively limited, but the recent break-through in neuroscientific technologies has given rise to a new interdisciplinary science that synthesizes the fields: neuroeconomics.

Neuroeconomics seeks to identify and understand the neural processes that underlie human decision-making by studying how the brain interacts with its environment to produce economic behavior. As such neuroeconomics is a crossroads between economics, psychology, and neuroscience that aims to better understand choice theory by unifying the separate approaches. In neuroeconomics theories and models are constrained by facts and by biological processes that determine how the brain functions. If a certain economic theory seems to describe choice behavior very accurately, but there is no evidence that the brain uses that model (for example due to limited computational power), then that theory can be discarded as a model of human behavior. Hence neuroeconomics tries more than just to *describe* choice behavior—it seeks to *understand* how decision-making works on the neural and cognitive levels.

The study of decision-making in neuroeconomics is vast, and it incorporates various subjects such as theory of choice under uncertainty, temporal discounting, framing effects, strategic choice, decision-making with respect to others, theory of mind, etc. (Glimcher 2003; McCabe 2003; Glimcher and Rustichini 2004; Sanfey, Loewenstein et al. 2006).

The insights that neuroeconomics can provide are best illustrated by an example: the Ellsberg paradox (Keynes 1921; Ellsberg 1961). Imagine 2 urns containing 100 balls each. Urn A (risky urn) contains exactly 50 black and 50 red balls, and Urn B (ambiguous urn) contains 100 black or red balls, the exact composition of which is unknown (Fig. 1). The balls are well mixed so that each ball is equally likely to be drawn. A bet on any color gives a payoff of \$20 if a ball of the chosen color is drawn, and \$0 otherwise.

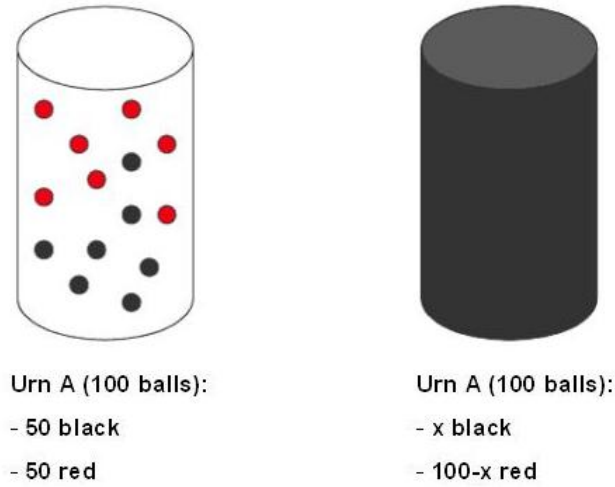


Fig. 1: Available choices in the Ellsberg paradox

In a situation like this most people prefer drawing a ball from Urn A rather than Urn B independently of the color, preferring known probabilities over unknown probabilities. If preferences are strictly based on probabilities, this pattern is a paradox leading to a violation of standard decision theory. Indeed, preferring a bet on a red ball from Urn A over Urn B implies that $p_{risk}(red) > p_{amb}(red)$. Similarly, preferring a bet on a blue ball from Urn A over Urn B implies that $p_{risk}(blue) > p_{amb}(blue)$. By adding together those two inequalities and taking into account the fact that the probabilities of red and blue must add up to 1, this leads to a contradiction:

$$\underbrace{p_{risk}(red) + p_{risk}(blue)}_{=1} > \underbrace{p_{amb}(red) + p_{amb}(blue)}_{=1}$$

This paradox can be resolved if one allows for subjective probabilities where $p_{amb}(red) + p_{amb}(blue) < 1$, e.g. $p_{amb}(red) = p_{amb}(blue) = 0.45$, the remaining 0.1 representing the amount of uncertainty that is associated with the ambiguous gamble. Although this trick solves the paradox mathematically, it does not explain how the brain makes a decision. Recently neuroeconomists investigated this paradox (Hsu, Bhatt et al. 2005; Huettel, Stowe et al. 2006), and found that different brain mechanisms are recruited to process risk and ambiguity. More specifically, they found that the amygdala and the lateral orbitofrontal cortex are more activated when facing ambiguous choices whereas the dorsal striatum is more activated in risky decision-making. They proposed a neural circuitry that responds to various degrees of uncertainty, in contrast to classical decision theory that makes no distinction between ambiguity and risk.

By investigating classical concepts and theories from behavioral economics (such as the Ellsberg paradox) and through the use of neuroscientific tools in combination with neuroscientific expertise, neuroeconomists try to create a mathematically and biologically plausible model of human behavior. Despite being a relatively new science, neuroeconomics has already significantly contributed to knowledge in a variety of areas. But it has also drawn some criticism, most notably from Gul and Pesendorfer who argue that neuroscience addresses different questions, and can therefore not provide any insight into economic theories (Gul and Pesendorfer 2005).

“Neuroscience evidence cannot refute economic models because the latter make no assumptions and draw no conclusions about the physiology of the brain. Conversely, brain science cannot revolutionize economics because the latter has no vehicle for addressing the concerns of economics.”

But one of the central assumptions of neuroeconomics is that more realistic models of human decision-making will lead to more accurate prediction of economic choice. Moreover, Gul and Pesendorfer argue that neuroeconomics addresses irrelevant questions,

as it focuses on what provides the most hedonic utility to subjects rather than the economic data itself:

“What makes individuals happy (‘true utility’) differs from what they choose. Economic welfare analysis should use true utility rather than the utilities governing choice (‘choice utility’).”

A lot of the criticism is also targeted towards neuromarketing, a closely related field that studies consumers’ brain responses to marketing stimuli (e.g., brand names, price, design) in order to provide better products and more efficient marketing campaigns. Most of that criticism comes from the popular press as well as from consumer protection agencies that are afraid that neuroeconomics and neuromarketing could be used to manipulate consumers’ choices. This is however far from the current standing of things as neuroeconomists are currently just trying to understand consumers’ choices.

Yet many behavioral economists, psychologists, and neuroscientists view neuroeconomics as a means to better understand how neural activity gives rise to a cognitive capacity for economic decision-making. Vernon Smith, 2002 Nobel Laureate in Economics, best expresses this optimistic attitude in his Nobel lecture "Constructivist and Ecological Rationality in Economics."

"New brain imaging technologies have motivated neuroeconomic studies of the internal order of the mind and its links with the spectrum of human decisions ... its promise suggests a fundamental change in how we think, observe, and model decision in all its contexts."

I.2. Experiments in Neuroeconomics

Neuroeconomics draws from related fields by using their tools and concepts in order to understand choice behavior. While behavioral economics and cognitive psychology provide the conceptual and mathematical frameworks as well as the experimental designs, neuroscience provides the scientific tools to study the neural correlates of choice behavior.

Neuroeconomics extends the approach used in behavioral economics by recording neural data (and also often psychophysiological data such as heart beat or skin conductance) in addition to behavioral data. In a typical behavioral economic experiment subjects are asked to choose between different options: for example, different gambles. By varying the experiment's parameters over a whole range of values (e.g., gambles with different stakes), economists create a model that predicts subjects' choices. Neuroeconomists check the biological validity of that model and try to improve its accuracy by making use of the neural data.

Experiments can range from very simple choice tasks to more sophisticated paradigms where subjects have to figure out non-trivial problems. Neuroeconomists are also interested in how people make choices with respect to others, and thus they conduct multi-subject experiments with the neural responses of one or more subjects being recorded simultaneously. The incentive for a subject is always to maximize the amount of money he can earn. Most experiments in neuroeconomics have been thoroughly studied by behavioral psychologists and/or economists (e.g., in game theory), which allows them to test the validity of existing models on neural data.

There are a variety of methods used in neuroscience to study neural activity, and each one has its own advantages and drawbacks, the most important feature being the spatiotemporal resolution (Fig. 2). Four of those methods are most often used in neuroeconomics: 1. imaging techniques (fMRI/PET), 2. electrophysiologic recordings, 3. lesions, and 4. drug manipulations.

Functional magnetic resonance imaging (fMRI) is the most commonly used method in neuroeconomics as it provides a great tradeoff between temporal and spatial resolution (down to 1–2 mm² and 0.5 sec). Furthermore it is non-invasive and does not require any tagging of the blood with radioisotopes (as does PET, for example). One of its drawbacks is that fMRI is only an indirect measure of brain activity (see Chapter 2 for more details). Better spatial and temporal resolutions are provided by electrophysiological measures such as single unit recordings or patch clamp techniques which record brain activity directly from neurons. However, because of its invasive nature, it is only used on animals.

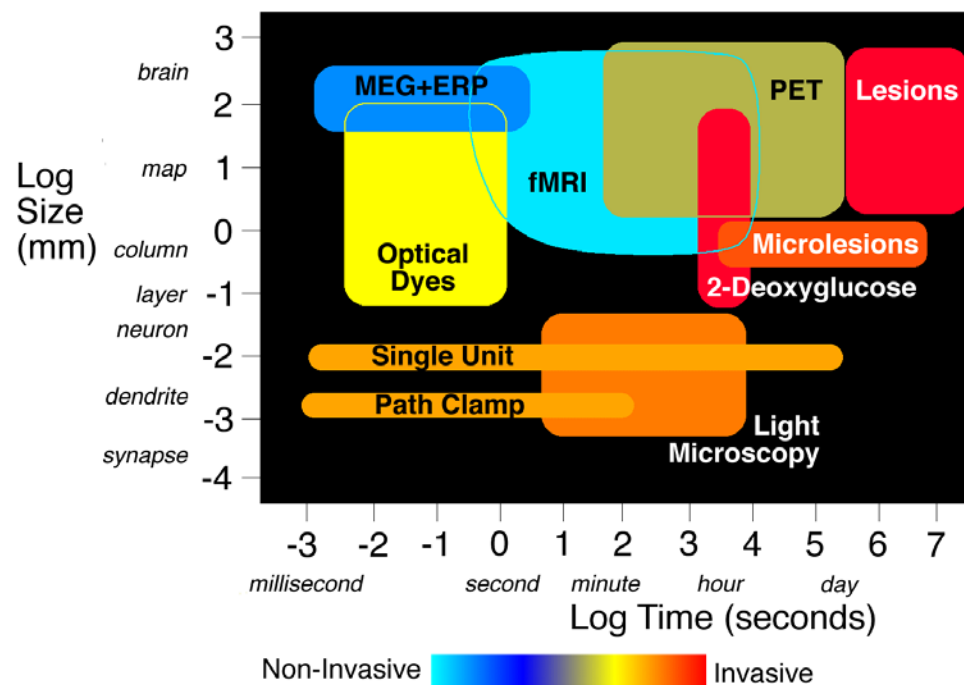


Fig. 2: **Spatiotemporal resolution of brain recording techniques** (from Jezzard et al., 2001). MEG = Magneto-EncephaloGraphy, ERP = Evoked Response Potentials, fMRI = functional Magnetic Resonance Imaging, PET = Positron Emission Tomography

The power of imaging methods is greatly improved when they are combined with other methods, and, in particular, lesion studies. If fMRI results have determined a specific area of the brain to be activated under a certain condition, then the same experiment can be administered to patients with localized lesions in that area. Such tests can assess whether that brain area is crucial to the correct execution of the task, or whether it can possibly be

compensated for by other brain structures in the complex neural circuitry. The number of lesions studies is however limited by the number of lesions patients, although brain lesions can be artificially created in animals. Finally, psychopharmacologic drugs can also be used to stimulate the brain and produce behavioral changes. For example, in a recent study it was shown that increasing the level of oxytocin in the body increases the level of trustworthiness of subjects (Kosfeld, Heinrichs et al. 2005). The drawback of such methods is that sometimes it is difficult to bypass the tough ethics standards imposed on researchers in administering drugs to subjects, and it is usually also difficult to assess how exactly the drug alters the normal functioning of the nervous system.

Chapter 2

BASICS OF FMRI

The results presented in this thesis have been obtained using functional magnetic resonance imaging (fMRI). In this chapter I will discuss how fMRI works and how fMRI data analysis is done. This chapter intends in no way to be complete, but only serves as an outline to describe the basic principles of fMRI required to understand the neuronal data analysis from neuroeconomic experiments. For a detailed review of fMRI methods see Jezzard or Huettel (Jezzard 2001; Huettel, Song et al. 2004).

II.1. MRI Physics

Magnetic Resonance Imaging (MRI) exploits the relaxation properties of the spin of atomic nuclei, and in particular the hydrogen nuclei. The spin is the angular momentum intrinsic to nuclei. In free space the spin of hydrogen nuclei are oriented randomly, but when a constant magnetic field \vec{B}_0 is applied, the spins will align either with or against the magnetic field (Fig. 3). There is a very small tendency for the spins to align parallel to the magnetic field (the difference is about one in a million), but due to the large quantity of nuclei in a small element of volume, this produces a detectable magnetic field M .

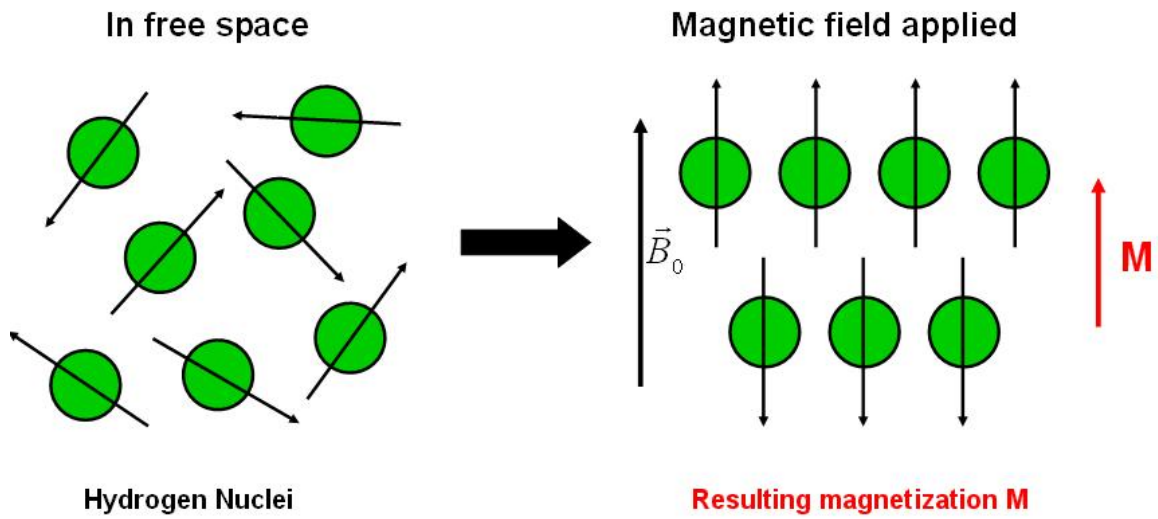


Fig. 3: Hydrogen nuclei in a magnetic field

When one applies a brief magnetic field orthogonal to \vec{B}_0 (called a 90 degree excitatory radio frequency or RF pulse), the aligned spins tip over to the transverse field. When the orthogonal magnetic field is removed, the spins do not immediately realign back with \vec{B}_0 , but they precess around \vec{B}_0 at a frequency directly proportional to it (Fig. 4) according to:

$$\omega_0 = \gamma B_0$$

where ω_0 is the precession frequency (also called Larmor frequency) and γ is the gyromagnetic ratio. This precession gives rise to a longitudinal magnetization M_l and a transverse magnetization M_t . As the spins gradually align back with \vec{B}_0 , the longitudinal magnetization grows back to M_0 , whereas the transverse magnetization decreases to 0. The times that it takes for these events to occur are known as the relaxation times T_1 (longitudinal) and T_2 (transverse), and they are an important aspect of MR imaging (Fig. 4). Surprisingly, T_1 and T_2 are not the same, and they are responsible for obtaining contrasts in magnetic resonance images.

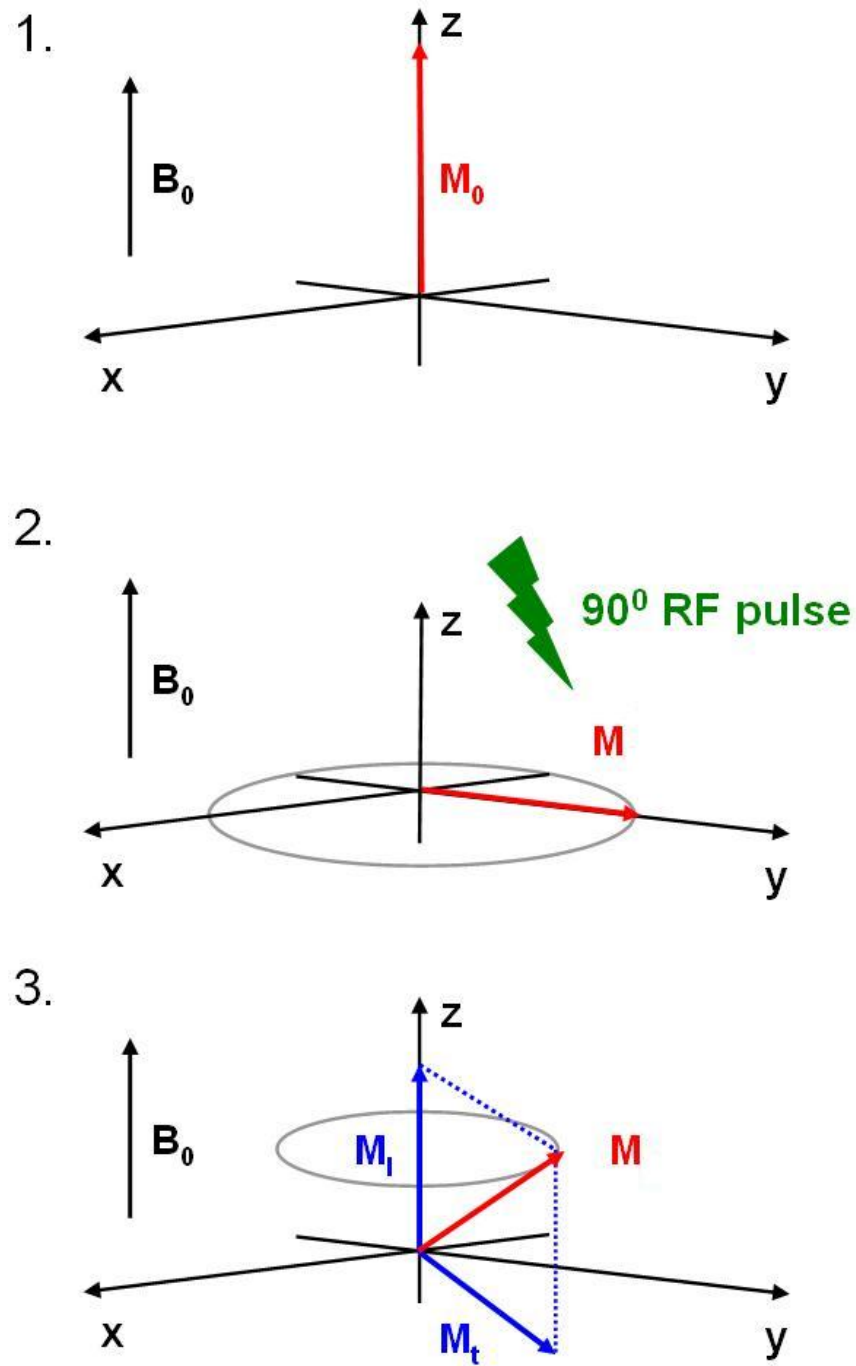


Fig. 4: **Effect of a RF pulse on the net magnetization.** 1. Net magnetization M_0 under the influence of a magnetic field B_0 . 2. A 90 degree RF pulse is applied, and tips the net magnetization over to the transverse field. 3. When the RF pulse is removed, the spins precess around B_0 , creating a net magnetization vector M that rotates around z , and gradually regains its original intensity M_0 . M can be decomposed into a longitudinal magnetization M_l and a transverse magnetization M_t .

The time course where M grows back to M_0 in the longitudinal direction is mathematically described by an exponential curve:

$$M_t = M_0(1 - e^{-t/T_1})$$

where t is time and T_1 depends on the nature of the tissue. The time it takes for the transverse magnetization to completely disappear is much shorter than T_1 . This is caused by small fluctuations in the precessing speed of the spins that cause them to gradually fall out of phase. T_2 is thus solely caused by spin-spin interactions and is independent of the nature of the tissue. This time course of such a relaxation is also described by an exponential curve:

$$M_t = M_0 e^{-t/T_2}$$

In reality this decay is actually much faster, because in addition to the spins interacting with each other, they are also affected by small inconsistencies in the applied magnetic field B_0 . When objects are placed in a magnetic field, they become magnetized themselves and changes in local magnetic susceptibility create distortions in the magnetic field. This results in a much faster decay of the transverse magnetization with a time constant T_2^* , which is dependent on the nature of the tissue.

T_1 -weighted MR images are obtained by measuring the decay times in different tissues of the brain. The connections of white matter have a long T_1 and appear white, whereas the congregations of neurons of gray matter have a short T_1 and appear gray. T_1 -weighted MR imaging is often used to obtain high-resolution anatomical images of the brain (Fig. 5).

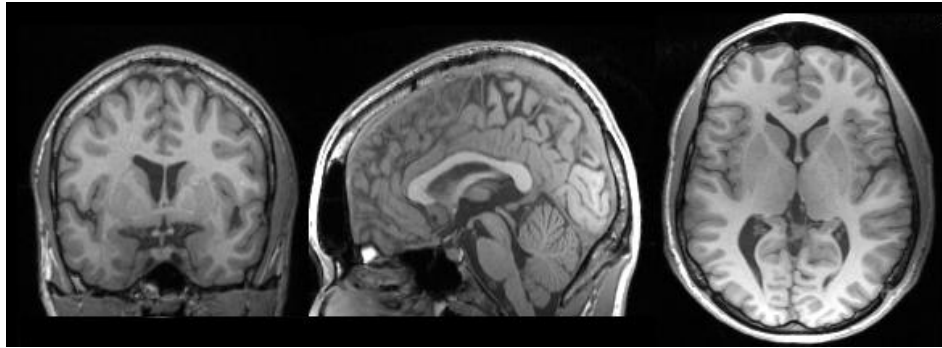


Fig. 5: **T1-weighted high-resolution anatomical MR image.**
From left to right: coronal view, sagittal view, axial view

II.2. Measuring the BOLD Signal

T_1 -weighted images allow us to discern between different tissues in the brain, but cannot provide any information about whether or not a brain structure is active during a certain task. This is achieved by using the fact that an increase/decrease in blood flow and blood oxygenation (known as hemodynamics) is a result of increased/decreased neuronal activity: firing neurons consume more of the oxygen that is being carried by hemoglobin in red blood cells than non-firing neurons consume. Oxygenated hemoglobin is diamagnetic and has the same magnetic properties as the rest of the tissue, whereas deoxygenated hemoglobin is paramagnetic. Deoxygenated hemoglobin causes a change in the magnetic susceptibility of the local blood supply, thereby causing an inhomogeneity in the local magnetic field, which in turn decreases the time constant T_2^* . This effect was first discovered by Ogawa (Ogawa, Lee et al. 1990), who showed that when mice breathed different concentrations of oxygen, the low concentration oxygen caused a significant signal drop in blood vessels in the brain. By measuring the time constant T_2^* , one can thus obtain Blood Oxygenation Level Dependent (BOLD) images which are the type of images that are acquired by fMRI imaging (Fig. 6). In T_2^* -weighted images white matter appears

grey and grey matter appears white, but the more interesting fact is that intensity changes as a function of brain activity.

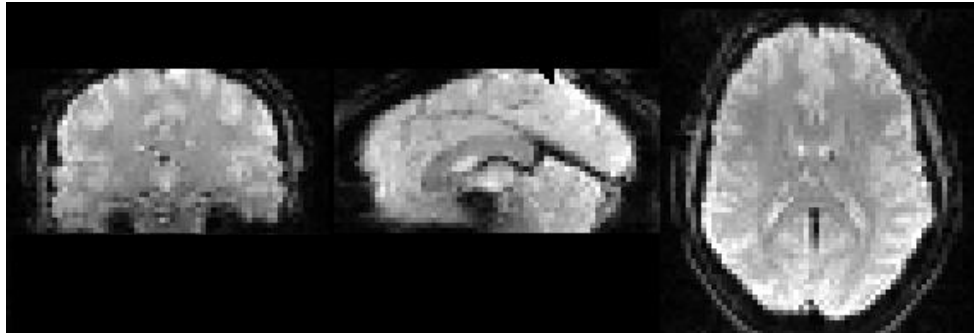


Fig. 6: T_2^* -weighted functional MR image. From left to right: coronal view, sagittal view, axial view

Although the connection between changes in blood flow and changes in neural activity has been known for a long time (Roy and Sherrington 1890), the exact cause of this relationship is still unclear. It has been argued that the increased blood flow fills the demand of oxygen and glucose needed for the restoration of energy supply during neuronal activity. This is consistent with the fact that the BOLD signal correlates most strongly with measures of presynaptic activity as shown by Logothetis (Logothetis, Pauls et al. 2001). Heeger has also shown that fMRI responses in the monkey medial temporal lobe correlate with single neuron firing rates (Heeger, Huk et al. 2000), and recently Logothetis's group has shown that negative fMRI responses correlate with decreases in neuronal activity in the monkey visual area V1 (Shmuel, Augath et al. 2006). Although these studies and other ones do not explain the reason for the relation between blood flow and neuronal firing, they do provide good evidence to use fMRI as a measure of neuronal activity.

One of the drawbacks of using the BOLD signal to measure neuronal activity is the slow hemodynamic response (~ 20 seconds, compared to the neuronal spiking which is on the order of milliseconds). The BOLD response has a characteristic shape (called hemodynamic response function or HRF): after an initial dip it peaks about 5–6 seconds after the onset, and then decays back to baseline after a small undershoot (Fig. 7). This

temporal limitation, in combination with the low spatial resolution, makes it impossible to record from individual neurons, but measures the activity of rather large sets of neurons.

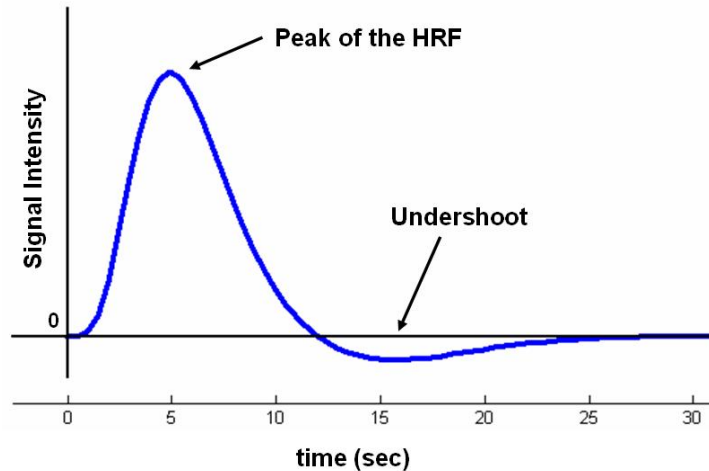


Fig. 7: Shape of the hemodynamic response function. Initial dip not shown here

II.3. fMRI Data Acquisition

In order to create 3-dimensional images of the brain, it is necessary to selectively measure the MR signal from individual volume elements in the brain, called voxels. This is achieved by applying three mutually orthogonal magnetic gradients in addition to the uniform magnetic field \vec{B}_0 . In the z-dimension this is achieved through a method called slice selection. Individual slices are selected by turning on a gradient during the excitatory RF pulse that tips the spins into the transverse plane. Since the spins are only tipped over at the Larmor frequency, the addition of the gradient assures that only one slice is being stimulated at a time. This process is repeated consecutively for all slices. Spatial encoding in each of the 2-D slices is done through frequency and phase encoding. Frequency encoding is achieved by turning on a magnetic gradient that changes the precessing frequency of the spins depending on their location along the x-axis. In the y dimension

another gradient is applied that causes the spins to be out of phase with respect to each other in a predictable manner along the y-axis. Both the frequency and phase encoding are then recovered through Fourier transform to recover the signal from a single voxel.



Fig. 8: Full-body human 3-T MRI scanner at Caltech

In a typical fMRI experiment subjects lie on their back in a MRI scanner (Fig. 8) with their head constrained by pads to avoid motion artifacts. They wear goggles that allow them to view a projected computer screen, and they hold response boxes in their hands to make choices during the experiment. Experiments last anywhere from 20 minutes to 2 hours, during which 3 sets of images are acquired: 1. a low resolution localizer that shows where the subject's head is positioned, 2. a high-resolution 3-D anatomical image that gives a detailed view of the structure of the brain (also called anatomical T_1 image), and 3. an fMRI image set that is composed of a series of 3-D pictures taken every 1–2 seconds (also called fMRI time-series).

Once the images have been acquired, the next step is to analyze them in order to determine which parts of the brain have been activated and what the nature of this activation is. Although this might seem to be a relatively simple problem, the task is made substantially more difficult by the fact that MR imaging introduces a lot of noise into the data. The most

important noise sources are thermal noise from the scanner, scanner drift, subject head motion, inhomogeneities in the magnetic field resulting from different magnetic susceptibility (e.g., the nasal cavity), physiological artifacts (e.g., respiratory cycle) and anatomical differences between subjects. Some of these sources of noise can be compensated for through optimizing scanning parameters and by image preprocessing, but the most powerful method is to repeat all stimuli over many trials and many subjects to allow for a statistical analysis. In the next two parts I will describe how the functional time-series and the anatomical image will be combined into a statistical activation map of the brain.

II.4. fMRI Data Preprocessing

The purpose of the preprocessing is twofold: it removes some of the introduced noise, and it prepares the fMRI images for a statistical analysis. There are five steps that need to be performed in the order listed below. All fMRI data preprocessing in this thesis was done using the statistical package SPM2 (Wellcome Department of Cognitive Neurology, London, UK, <http://www.fil.ion.ucl.ac.uk/spm/software/spm2>).

Slice Time Correction

When acquiring a fMRI time-series, a complete snapshot of the brain is taken every $TR=2$ sec (TR = repetition time). However, these snapshots are not taken instantaneously, but consecutively for every slice in the brain. Hence, between the first and the last acquired slice of the brain, the time difference is close to 2 seconds. In addition to that the slices are not acquired in order, but in an interleaved way to minimize noise introduced from adjacent slices. To compensate for these effects, one takes advantage of the fact that each slice is repeatedly acquired every TR , and uses a sinc filter to interpolate all slices to a reference slice (e.g., the middle slice).

Realignment

One of major sources of noise in fMRI data is subject head movement. Although head restraints in the MRI scanner limit head motion to a minimum (typically less than 3 mm, i.e., the size of a voxel), the remaining motion artifacts need to be corrected for by using a realignment process. Since it can be safely assumed that the dimensions of the brain are not changing over the course of the experiment (no scaling or shearing), this can be done using a rigid body transformation with 3 rotation and 3 translation parameters (Friston, Williams et al. 1996), described by the sequence of 4 matrices below. All images are realigned sequentially with respect to the previous image so that at the end all images are realigned with respect to the first image.

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & X_{trans} \\ 0 & 1 & 0 & Y_{trans} \\ 0 & 0 & 1 & Z_{trans} \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{translation}} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\Phi & \sin\Phi & 0 \\ 0 & -\sin\Phi & \cos\Phi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{x\text{-pitch}} \underbrace{\begin{pmatrix} \cos\Theta & 0 & \sin\Theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\Theta & 0 & \cos\Theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{y\text{-roll}} \underbrace{\begin{pmatrix} \cos\Omega & \sin\Omega & 0 & 0 \\ -\sin\Omega & \cos\Omega & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{z\text{-yaw}}$$

Coregistration

As functional MR images measure brain activity and are acquired every TR resulting in poor spatial resolution, they are a poor indicator of the underlying brain structure. Since it is not only important to determine how the brain is activated but also where it is activated, the high-resolution T_1 image is used to map brain activations to brain areas. Similarly as for the realignment the T_1 image and functional images are coregistered using a 12-parameter affine transformation (see equation below). This can be done in several different ways by finding the optimal parameters that minimize/maximize a cost function (e.g., mean squared difference, normalized cross-correlation, or normalized mutual information between the two images).

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \\ z_0 \\ 1 \end{pmatrix}$$

where (x_0, y_0, z_0) and (x_1, y_1, z_1) are the coordinates before and after the coregistration, respectively (Frackowiak, Ashburner et al. 2002).

Normalization

If one wants to generalize results about brain function, the variability in brain structure across subjects needs to be taken into account. There are two ways this can be done. One can limit the analysis to a predetermined region of interest, and then compare brain activations only in that region. This can however only be done for studies with a priori hypothesis about involved the brain structures. Another problem is that the structure in question needs be easily identifiable. A much more common solution is to normalize each subject's anatomical MR image to a canonical average brain (Fig. 9). The most commonly used one is the MNI-template from the Montreal Neurological Institute (Evans, Collins et al. 1993), which is an average of 305 anatomical MR images. Normalization is done using a combination of linear and non-linear warping functions (Ashburner and Friston 1999).

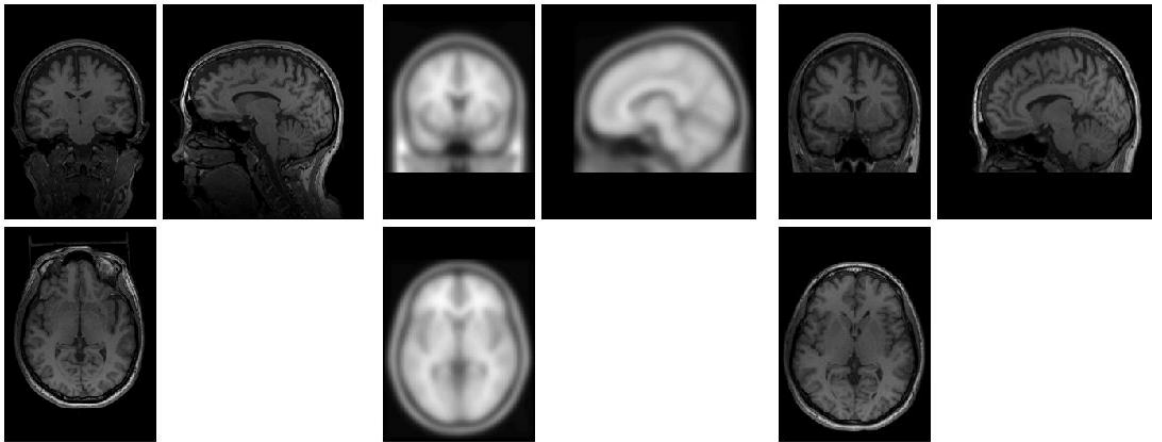


Fig. 9: **Normalization procedure.** Leftmost panels: Original anatomical brain. Center panels: Template or average brain. Rightmost panels: same anatomical brain after normalization

Smoothing

The signal to noise ratio in fMRI data is typically very low, at the order of 1%. To improve the quality of the data, both temporal and spatial smoothing is performed. Temporal smoothing is achieved by filtering each voxel's time-course with a low pass-filter. Spatial smoothing is achieved by filtering each image with a three-dimensional Gaussian smoothing kernel with FWHM=8 mm (Full Width at Half Maximum), i.e., 2–3 times the size of a voxel (Fig. 10).

Even before the spatial smoothing is performed, neighboring voxels are correlated because of the way fMRI data is acquired. It is very difficult to estimate these correlations, but by filtering the images spatially with the Gaussian kernel, stronger and known correlations are imposed onto the data. This turns out to be very useful for some parts of the subsequent statistical analysis.

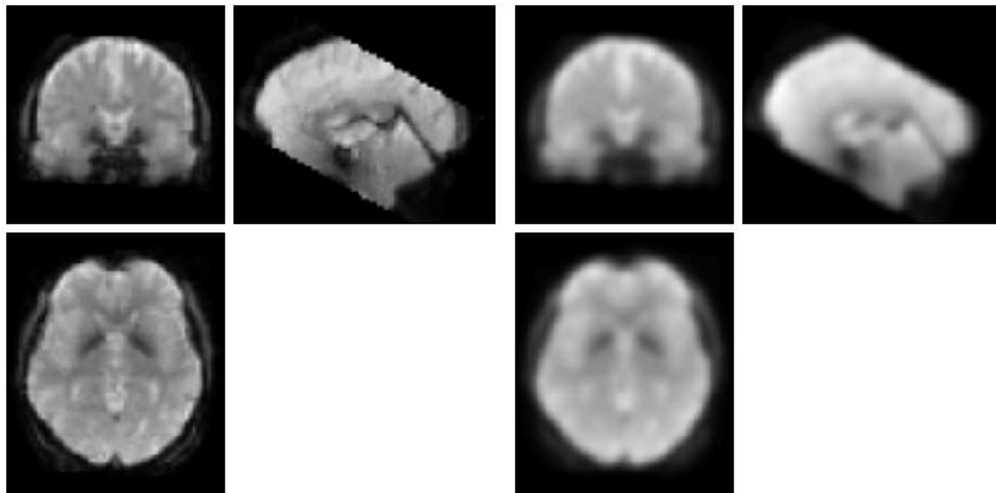


Fig. 10: **Spatial smoothing procedure.** Functional imaging data before (left panels) and after (right panels) spatial smoothing with an 8 mm FWHM Gaussian kernel

II.5. fMRI Data Analysis

After the preprocessing the fMRI data is ready for further analysis, but it still has a low signal-to-noise ratio (Fig. 11). Hence it is impossible to detect individual events, and one needs to perform a statistical analysis. In the following I will describe the most commonly used method to analyze fMRI data, namely the general linear model (GLM). All of the GLM analysis in this thesis has been done using the statistical package SPM2.

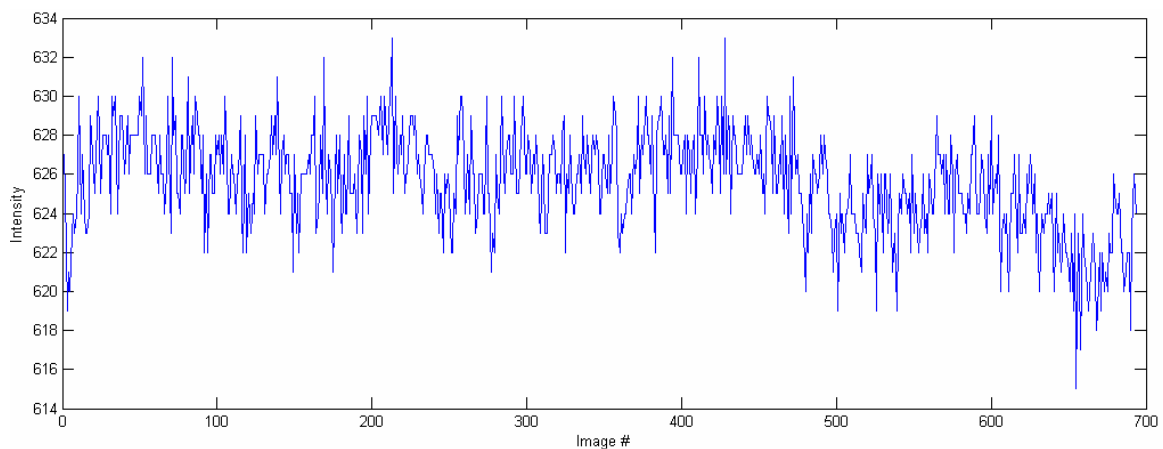


Fig. 11: Sample time-series of a voxel within the brain

The main idea of the GLM is to obtain statistics about how well a series of observations (the fMRI data) can be described by a linear combination of explanatory variables (the stimuli and/or subject responses). This requires the experimenter to have an a priori hypothesis about the time and shape of the brain response, but not about the location, as the analysis is done on a voxel-by-voxel basis over the whole brain. In the following, the GLM method is described with respect to an individual voxel, but the same method is applied to all voxels.

Suppose that during an fMRI experiment we acquire N images. Now for a given voxel we can create a time-series $y = (y_1, y_2, \dots, y_N)$ where y_i represents the intensity of that voxel at time-step i . This is considered to be the independent variable. Now let's assume that we

also have 2 explanatory variables $x_1 = (x_{11}, x_{12}, \dots, x_{1N})$ and $x_2 = (x_{21}, x_{22}, \dots, x_{2N})$ that could be used to describe the data, and we are interested in determining the linear fit between the data and the explanatory variables. We can then write the independent variable y as a linear combination of x_1 and x_2 (also called regressors) plus a constant term and an error term:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 x_{1N} + \beta_2 x_{2N} + \varepsilon_N \end{aligned}$$

where β_1 and β_2 are the unknown parameters describing the relation between y and x_1 and x_2 , and where β_0 is a constant term. The errors ε_i are independent and identically distributed normal variables with zero mean and variance σ^2 , i.e., $\varepsilon_i \sim N(0, \sigma^2)$. This relation can be written in matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

or equivalently:

$$y = X\beta + \varepsilon \quad (1).$$

Note that under this form there can be an unspecified number of explanatory variables in X . But before solving this equation, one needs to take into account the fact that the BOLD response in the brain to a punctuate stimulus is not punctuate. Indeed, as described in Section II.2. and Fig.7, the BOLD response of the brain is a hemodynamic response function (hrf). The shape of this hrf has been determined experimentally (Blamire, Ogawa et al. 1992; Buckner, Bandettini et al. 1996), and it has been shown to vary slightly across

brain regions (Buckner, Bandettini et al. 1996; Ollinger, Shulman et al. 2001). Friston et al. have shown that the hrf can be approximated by the sum of two gamma functions, one modeling the peak and one modeling the undershoot (Friston, Fletcher et al. 1998; Frackowiak, Ashburner et al. 2002):

$$hrf(\tau) = \left(\frac{\tau - o_1}{d_1} \right)^{p_1-1} \frac{e^{-(\tau-o_1)/d_1}}{d_1(p_1-1)!} + \left(\frac{\tau - o_2}{d_2} \right)^{p_2-1} \frac{e^{-(\tau-o_2)/d_2}}{d_2(p_2-1)!}$$

where o_i is the onset delay, d_i is the time-scaling and p_i is an integer phase-delay ($i=1,2$). Before solving (1), each column of the matrix X (other than the first column) is thus convolved with $hrf(\tau)$ to give a new matrix \tilde{X} , also called the design matrix:

$$\tilde{X} = X \otimes hrf(\tau) = (\underline{1} \quad x_1 \otimes hrf(\tau) \quad x_2 \otimes hrf(\tau) \quad \dots \quad x_N \otimes hrf(\tau))$$

where $\underline{1}$ and x_i are the column vectors of X . The effects of this convolution are illustrated in Fig. 12.

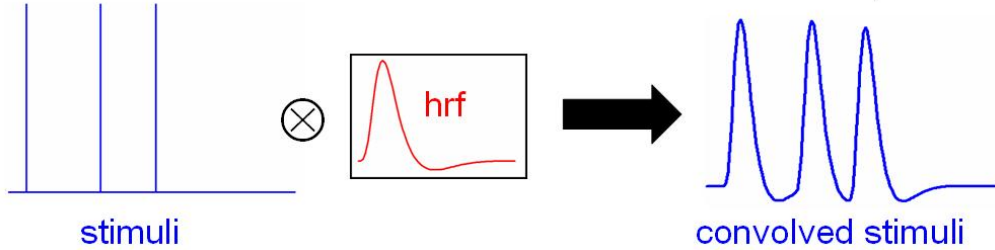


Fig. 12: **Effect of the convolution with an hrf.** The spikes (left panel) represent punctuate stimuli that are convolved with an hrf (middle panel) to model the BOLD response (right panel).

Now (1) can be written as:

$$y = \tilde{X}\beta + \varepsilon$$

which can be solved using ordinary least squares to give the parameter estimates:

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y.$$

If \tilde{X} is of full rank (i.e., if the columns of \tilde{X} are linearly independent), it can be shown that the parameter estimates are uniformly distributed: $\hat{\beta} \sim N(\beta, \sigma^2 (\tilde{X}^T \tilde{X})^{-1})$. This result can now be used to determine if there is significant activation for a voxel with respect to one or more of the regressors. This is achieved through the use of t-tests between β values, a manipulation called contrast-estimates. The used t-statistic is:

$$\frac{c^T \hat{\beta} - c^T \beta}{\sqrt{\hat{\sigma}^2 c^T (\tilde{X}^T \tilde{X})^{-1} c}} \sim t_{N-p}$$

where t_{N-p} is a Student's t-distribution with $N - p$ degrees of freedom, and c is a contrast vector.

There are two main types of t-tests that can be performed: the first type tests for effects among regressors and the null hypothesis is $H_0 : c^T \beta = 0$. For example, in a design with 4 conditions, if one wants to assess whether a particular voxel was activated differently under condition 2 (regressor 2) than under condition 3 (regressor 3), the contrast vector c is: $c = (0 \ 0 \ 1 \ -1 \ 0)$, corresponding to $H_0 : c^T \beta = 0 \Leftrightarrow \beta_2 - \beta_3 = 0 \Leftrightarrow \beta_2 = \beta_3$. The second type tests for effects between a regressor and baseline (the constant term in \tilde{X} with the corresponding parameter estimate β_0). For example, in the same design if one wants to test whether a particular voxel was activated under condition 2, the contrast vector c is: $c = (0 \ 0 \ 1 \ 0 \ 0)^T$, corresponding to $H_0 : \beta_2 = 0$.

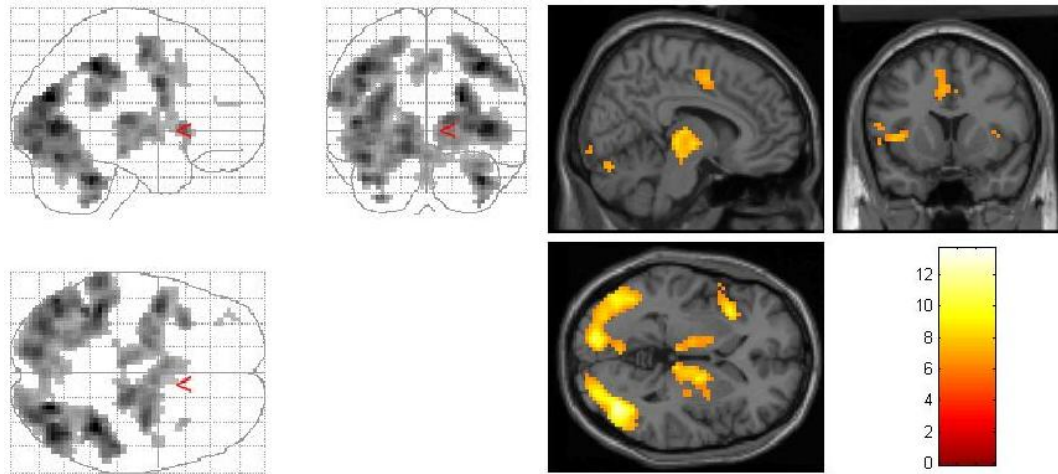


Fig. 13: **fMRI activations in glass-brain and on SPM map.** Two typical ways of displaying fMRI results: in a transparent glass brain in the left-hand panels, and the same activation on a color-coded SPM map in the right-hand panels. The legend on the right indicates the t-value.

These t-tests are performed on all voxels of the brain to give a statistical parametric map (SPM), which is color-coded and overlaid on the high-resolution anatomical scan to give the characteristic fMRI activation map (Fig. 13). This statistical analysis is often called a fixed effects analysis because it assumes that the subject's brain response to each single instance of a particular event is identical. But in most fMRI experiments one is interested in drawing conclusions that hold with respect to all subjects in the dataset. To do this, a random effect analysis is performed in which every subject is treated as an independent observation. This 2nd level is achieved by simply doing a t-test on the contrast values for all subjects (on a voxel-by-voxel basis).

It should be noted that there are many adjustments that have not been mentioned in this chapter, and which significantly improve the GML method, in particular with respect to spatial correlations, estimation of noise terms, false positives, filtering etc. Although the GLM is the most commonly used method to analyze fMRI data, it has some drawbacks, the most important one being the assumption of linearity. When performing a GLM analysis, one assumes that the BOLD response is linear, i.e. that the responses to several subsequent stimuli sum up linearly. There is however evidence that this is not the case, and that the

non-linearity increases as the stimuli are closer in time. Another major weakness of the GLM is that it is a model-driven analysis, i.e., the experimenter needs to have an a priori hypothesis about the variables that produce the neural activity (their shape through the hrf and their timing). Those issues can be addressed by other types of analyses (described in Chapter 3), most notably a region-of-interest (ROI) analysis that focuses on a particular brain area, as well as a newly developed correlational analysis (Section III.5).

NEURAL CORRELATES OF ECONOMIC DECISION-MAKING

This chapter analyzes various neural aspects of economic decision-making in the Trust Game, which is a 2-person social exchange game used in the field of behavioral game theory. The results are grouped into 5 sections:

- Behavioral Results (Section III.3)
- Strategic Uncertainty and Prediction Analysis (Section III.4)
- Dynamic Cross-Brain Analysis (Section III.5)
- Trust & Cooperation (Section III.6)
- Agency Attribution (Section III.7)

The chapter starts off by presenting some background information about behavioral game theory and about existing studies on social exchange in neuroeconomics.

III.1. Background

III.1.1. Behavioral Game Theory

Behavioral game theory is a branch of economics that studies situations (or games) in which players interact with each other through a series of decisions made to maximize their own returns (Camerer 2003). Each game consists of a set of players, a set of options (called

strategies) available to the players and a set of outcomes for each combination of strategies (known to the players). Hence the games are well-defined mathematically and can be studied in terms of optimal strategies, equilibriums, etc. In a typical game the players are assumed to always act rationally in order to maximize their earnings (according to the *homo economicus* model), which leads to a game theoretic solution. However, when people actually play these games, their behavior is often irrational from a game theory perspective, e.g., sometimes they make decisions to maximize the group's earnings (instead of their own). Although only a very small fraction of people play according to the predictions from game theory, the theory still provides a reasonably valid description of human behavior, and can be used as a model to predict how people ought to behave.

The ultimatum game (UG) is a game used in behavioral game theory to test how much people deviate from rational behavior (Fig. 14). In this game one player (the proposer) is asked to split a certain amount of money between himself and another anonymous player (the responder). If the responder rejects the offer, nobody gets anything; if he accepts the offer, the money is split according to the division the proposer proposed.

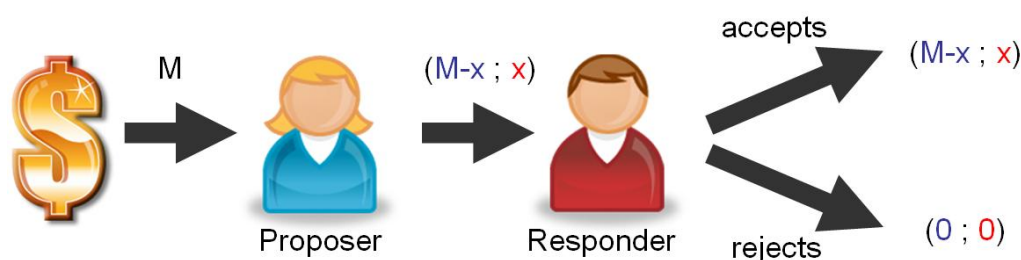





Fig. 14: The ultimatum game

From a game-theoretic perspective the responder should accept any non-zero offer (since rejecting will not give him any money). Hence the proposer should offer the smallest amount possible since he knows that the responder will accept any non-zero offer. However humans significantly diverge from this strategy: the median and mean offers are usually 40–50% and 30–40%. 40–50% offers are almost always accepted and offers below 20% are rejected half the time. There are two main explanations for this behavior:

responders are either fair-minded (or altruistic) or/and they are afraid that low offers will be rejected. The contribution of each one of those two factors can be captured by measuring the proposer's level of altruism in another game, the dictator game (DG). This game is the same as the UG, except for the fact that the second player has to accept any offer. Thus any non-zero offer in the DG by the proposer is purely altruistic, and can be used to explain that the proposer's offer in the UG is not just purely strategic.

Another famous game from behavioral game theory is the prisoner's dilemma (PD). Figure 15 shows the pay-off matrix in a typical PD experiment as well as its more general form. Mutual cooperation pays off $C=2$ for each player, which is better than mutual defection which only pays $D=1$ for each player. If one player defects and the other one cooperates, the defector earns $T=4$ which is better than the payoff from cooperation, whereas the cooperator earns $S=0$, which is less than the payoff from defection. Since $T=4 > C=2$ and $D=1 > S=0$, both players prefer to defect independently of whether the other player cooperates or defects. Hence the Nash equilibrium is mutual defection although it pays off less than mutual cooperation.

A. Example					
 		Cooperate		Defect	
 Cooperate		2	2	0	4
		4	0	1	1
Defect					




B. Generalized Form					
 		Cooperate		Defect	
 Cooperate		C	C	S	T
		T	S	D	D
Defect					

Fig. 15: **The prisoner's dilemma.** A. Example of payoff structure: first amount listed denotes row player's payoff, and second amount denotes column player's payoff. B. Generalized form of the Prisoner's Dilemma with the assumption: $T > C > D > S$

Another game that studies cooperation and defection is the public goods game (PG). In the PG game (Fig. 16), N players can invest a certain portion p_i of their initial endowment M into a common pot (the public good). The public good earns a return (it is multiplied by

a factor $f > 1$), which is then split evenly among all players such that player i receives a total of $M - p_i + \frac{f}{N} \sum_{j=1}^N p_j$. The optimal solution is to invest nothing, and to pick up other people's investments. If everyone cooperated however, the players would maximize their total collective earnings.

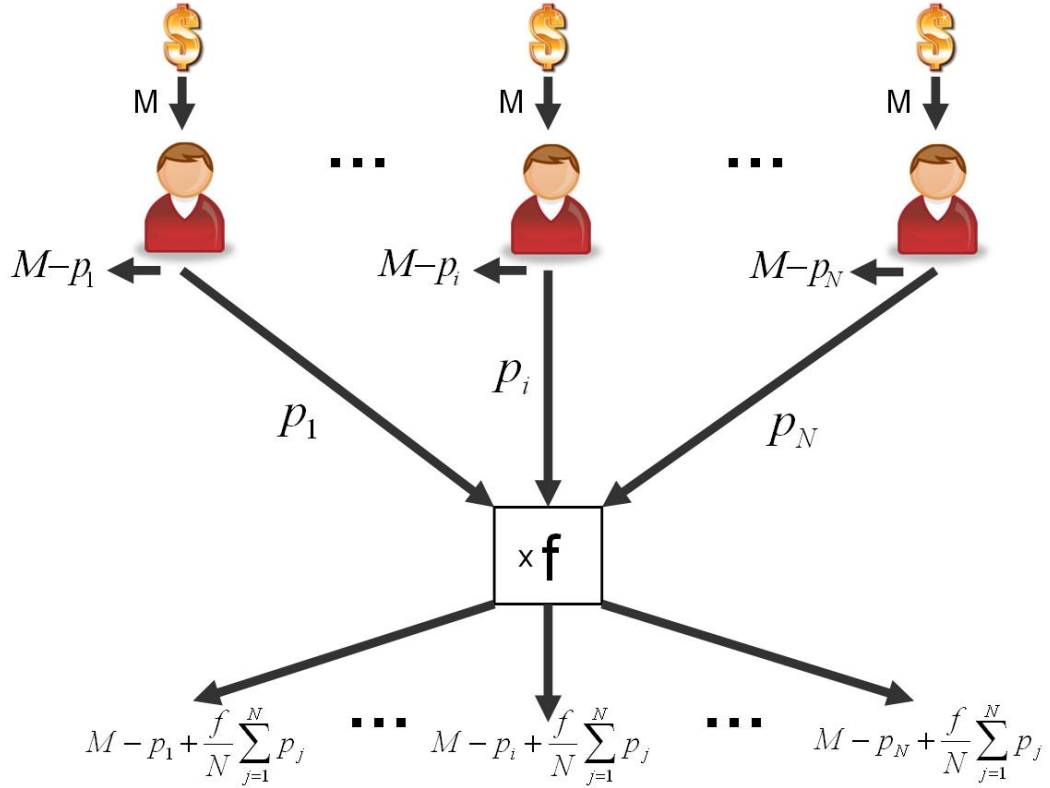


Fig. 16: The public goods game

For the purpose of this thesis another type of game was used: the trust game. It combines the notions of cooperation and defection of the PD and the principle of investing into a common good of the PG games with the sequential nature of the UG and DG. Also, the existence of a mathematical framework makes all of these games (and the trust game in particular) very appealing for study with neuroscientific tools.

III.1.2. The Trust Game

The trust game (Camerer and Weigelt 1988; Berg, Dickhaut et al. 1995) is a modified version of the dictator game (see Fig. 17): one player, the Investor, is endowed with M dollars and can invest any portion of it. The invested amount x is then multiplied by f , and the second player (the Trustee) decides how much of the resulting investment to keep and how much to pay back. The investor's payoff is the amount originally held back, $M-x$, plus the returned money y . The Trustee's payoff is $fx-y$. The collective gain is $(M-x+y)+(fx-y)=M+(f-1)x$ which is maximized for $x=M$ (when the Investor invests everything). Thus there is a significantly larger gain from cooperation.

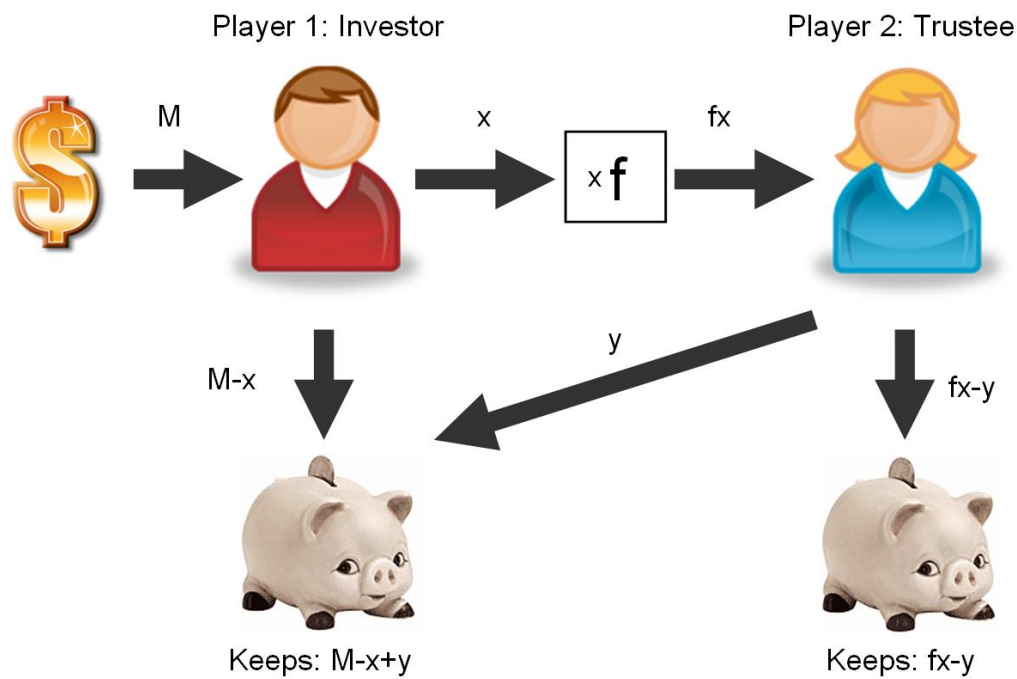


Fig. 17: The trust game

According to game theory the best strategy in this game is for the Investor to invest nothing. Indeed, since the Trustee is trying to maximize his own earnings, he will not return any money, and thus there is no incentive for the Investor to invest.

The trust game becomes significantly more interesting and complex when it is repeated with the same players over several rounds. It can be seen as a simplified model of social

interaction, where the multiplication of x represents productivity, the invested money amount x is a measure of trust, and the returned money amount y is a measure of trustworthiness. The repetitive nature of the trust game allows for multiple interactions between players, and thus more diverse and complex strategies. As players pass money between each other, they create (or break up) a mutual trust relationship based on reputation and previous history. The strength of that relationship can be measured in terms of reciprocity (this method will be explained in subsequent sections).

Although the repeated trust game is substantially more complex than its single-shot counterpart, the game-theoretic solution remains unchanged: the Investor should never invest money, and the Trustee should keep all the money that he receives. Indeed, the situation in the last round of the repeated trust game is exactly the same as it is in the single-round trust game, and neither player should invest/repay anything. Considering that nothing happens in the last round, the second-to-last round can now be considered to be the “last” round of the game, and the same reasoning applies. By reiterating this process backwards over all rounds, it follows that every round should be treated as a single-shot game, and that the Investor should never invest any money. For an initial endowment of $M = 20$, a multiplication factor $f = 3$, and 10 rounds, the overall earnings will be $10 \times 20 = 200$ (all earned by the Investor), which is considerably less than the maximum overall earnings from a cooperative strategy ($10 \times 20 \times 3 = 600$).

The invest-zero strategy is only played by a very small fraction of people, and most subject pairs cooperate up to some degree to maximize their earnings. The trust game has been studied by economists in many countries, for different monetary pay-offs and under various experimental conditions. The basic finding is that the social exchange in the first few rounds is based on reciprocity, i.e. both players invest and return increasingly more money. In later rounds Trustees return less money which leads the Investors to invest less money.

An important variable in the trust game is the notion of fairness. Although not all subjects share the same opinion about what is a fair or unfair split-up of the money, they all tend to heavily base their decisions on that metric. This is illustrated through one notion of fairness

(for $M = 20$, $f = 3$): a round is defined to be fair if the Investor and the Trustee both earn the same amount of money.

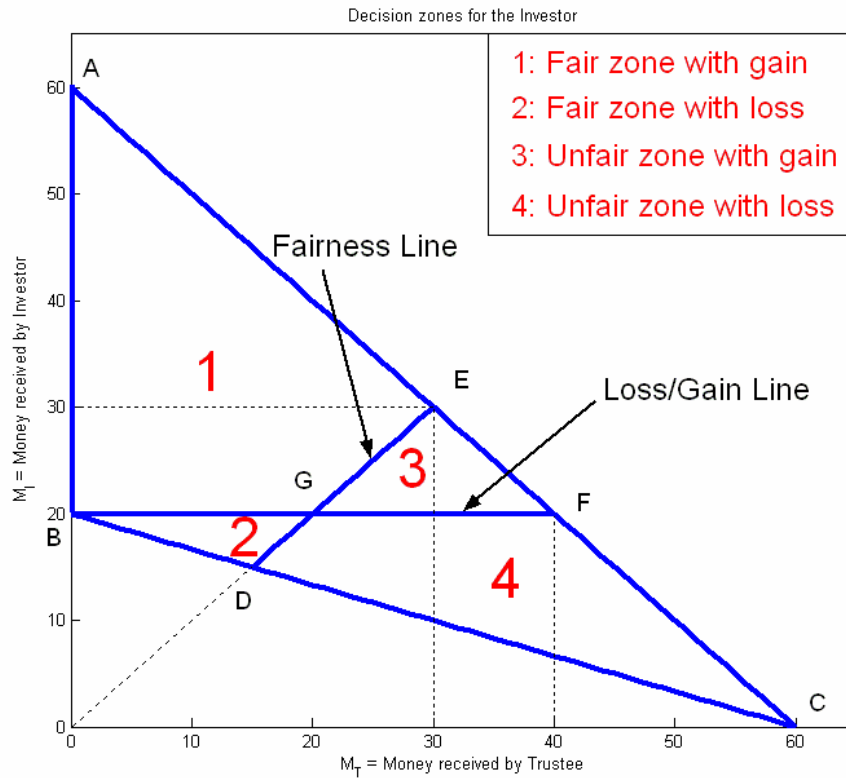


Fig. 18: Fairness in the trust game

If the Investor invests at least 5, the Trustee can always return enough money to make the round fair. In some cases the Investor can end up with less money than his initial endowment, even though the Trustee divided the money in a fair way. In other cases he ends up with more than his initial endowment although the round was unfair. Figure 18 illustrates this from the point of view of the Investor. All possible money split-ups (M_I, M_T) fall within the triangle ABC (where M_I is the money earned by the Investor and M_T is the money earned by the Trustee). Any money split-ups that fall on the segment $[D, E]$ are considered to be fair. Zone 1 (ABGE) is a hyper-fair zone where the Investor gets more money than the Trustee and more than his initial endowment of 20. The split-up

in Zone 2 (BDG), is also hyper-fair, but the Investor receives less than his initial endowment. Zones 3 (EFG) and 4 (CDGF) are both unfair, although in Zone 3 the Investor earns more money than his initial endowment (and thus still profits from the investment).

III.1.3. Previous Studies

All of the games mentioned in the previous sections have been studied extensively in behavioral economics (Camerer 2003), and with the recent availability of neuroscientific tools they become increasingly more popular in neuroeconomics.

McCabe et al. (McCabe, Houser et al. 2001) performed one of the earliest experiments in neuroeconomics by implementing a simplified version of the trust game. They found that subjects who played cooperative strategies have increased activity in the inferior frontal gyrus when playing against another person compared to when playing against a computer. In another implementation of the trust game, Delgado et al. (Delgado, Frank et al. 2005) modulated the investor's a priori perception of the trustee's moral character. They found that the caudate was differentially activated with respect to positive and negative feedback, but only when subjects were playing with the neutral partner. Kosfeld et al. (Kosfeld, Heinrichs et al. 2005) have also been able to artificially increase the level of trust in investors by intranasal administration of oxytocin, a neuropeptide that plays a key role in social attachment and affiliation.

The neural correlates of cooperative social behavior have been investigated in an iterated version of the Prisoner's dilemma (Rilling, Gutman et al. 2002), where it was shown that mutual cooperation was associated with activity in reward processing structures such as the nucleus accumbens, the caudate nucleus, the ventromedial frontal/orbitofrontal cortex and the anterior cingulate cortex. The prisoner's dilemma was also recently used to study the behavioral and emotional responses to conflict and cooperation in a special population group (McClure, Parrish et al. 2007). One of the main findings was adolescents with anxiety/depressive disorders responded more cooperatively to cooperative overtures from

their co-players, suggesting a first step towards understanding the mechanisms underlying social impairment. In another type of game Decety et al. (Decety, Jackson et al. 2004) studied the neural correlates of cooperation vs. competition, and found the orbitofrontal cortex to be more activated in the cooperation condition, and the inferior parietal and medial prefrontal cortices to be more activated in the competition condition.

Sanfey et al. studied the neural correlates of unfairness in a ultimatum game (Sanfey, Rilling et al. 2003), and found that receiving unfair offers activated brain areas related to both emotion (anterior insula) and cognition (dorsolateral prefrontal cortex). Moreover, the activity in the insula was correlated with subject's decision to reject the offer. In a more complicated design, subjects who received unfair offers were able to punish their partner by using some of their own money (de Quervain, Fischbacher et al. 2004). Punishments elicited activations in the dorsal striatum (a reward processing structure), and the level of activation was correlated with their willingness to incur greater loss in order to punish.

Although a couple of neuroimaging studies have investigated a simplified version of the trust game, the work presented in this thesis is the first to analyze the neural correlates of the repeated trust game in interacting subjects.

III.2. Experimental Design and Methods

III.2.1. Task

Subjects played an anonymous 10-round trust game. At the beginning of each round, the Investor received 20 monetary units (mu), and was able to invest any portion of it (in 1mu increments) with the other player (Trustee). The invested amount, denoted x , was then tripled and the Trustee decided how much of the resulting investment to pay back. At the beginning of each round the Investor received 20 new mu. Roles were fixed throughout the experiment. Earned mus were accumulated over rounds, and at the end of the experiment subjects were paid a monotonic step function of their actual experimental earnings: 0–67mu: \$20; 68–133mu: \$25; 134–200mu: \$30; 201–300mu: \$35; >300mu: \$40. Subjects

had no knowledge of the actual pay scale, but were informed that they could earn \$20–40 based on their performance.

III.2.2. Subjects

To assure anonymity, subjects were recruited from separate subject pools at the California Institute of Technology (CIT), Pasadena, CA and Baylor College of Medicine (BCM), Houston, TX. Informed consent was obtained by using a consent form approved by the Internal Review Boards of both CIT and BCM. Investor/Trustee roles were assigned pseudo-randomly, and subjects were matched for gender, location and player role to control for confounding effects. Specifically, there were 12 subject pairs of each combination MM, MF, FM, FF where M=Male player and F=Female player, the first subject listed denoting the Investor and the second one the Trustee. There were a total of 48 subject pairs.

III.2.3. Experimental Setup

The behavioral and functional data in the trust game was acquired using the NEMO hyperscanning software (Montague, Berns et al. 2002), which simultaneously recorded BOLD responses in interacting subjects (Fig. 19). Subjects were instructed identically, but separately at each location (instructors read a script describing the task while showing screenshots of the game).

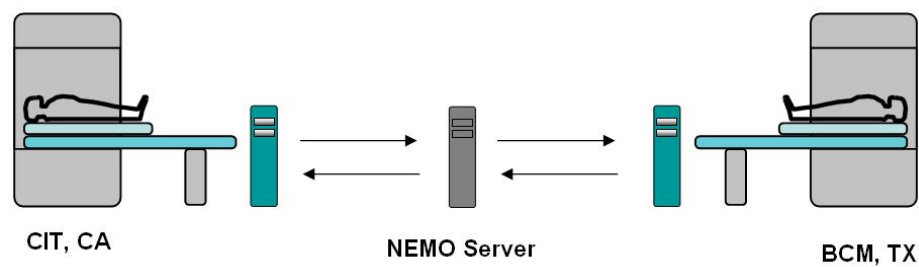


Fig. 19: Hardware setup of the trust game

Stimuli were presented through MRI compatible goggles (Resonance Technology) at CIT and a back-projected screen at BCM. Subjects used MRI-compatible button boxes to make their decisions by toggling a slider bar up and down.

The timeline of a single round of the trust game is depicted in Fig. 20. Each round starts with a blank screen that lasts 4 seconds, followed by a free response period where the Investor decides how much money to invest. During that period the Trustee sees a blank screen. 8 seconds after the Investor submits his decision, the results of the investment phase are revealed to both subjects simultaneously. Then the repayment phase starts where the Trustee decides how much to send back to the Investor and how much to keep. During that time the Investor sees a blank screen. 8 seconds after the Trustee's decision the results of the repayment phase are revealed simultaneously to both players. After another 8 second blank screen a summary with the overall totals for the round is revealed to both subjects. Each round is separated from the next one by a blank screen of random duration (12–42 seconds). Note that except for the periods of free response both players view the same visual stimulus.

III.2.4. fMRI Data Acquisition and Preprocessing

Brain image acquisition was done on a Siemens Trio (CIT) and a Siemens 3T Allegra (BCM). High resolution T1-weighted scans (0.48 mm x 0.48 mm x 1 mm) were acquired using a MPRage sequence. Functional images were acquired using echo-planar T2* images with BOLD contrast. Parameters were as follows: repetition time (TR) = 2000 ms; echo time (TE) = 40 ms; slice thickness = 4 mm yielding in a 64x64x26 matrix (3.4 mm x 3.4 mm x 4 mm); flip angle = 90 degrees; FOV read = 220 mm; FOV phase = 100 mm, series order: interleaved.

Imaging data was preprocessed using SPM2, and included slice time correction, motion correction, coregistration, normalization to the MNI template and smoothing of the functional data with an 8 mm kernel (see Section II.4 for details). During the preprocessing steps, all voxels within an image were resized to 3 mm x 3 mm x 3 mm.

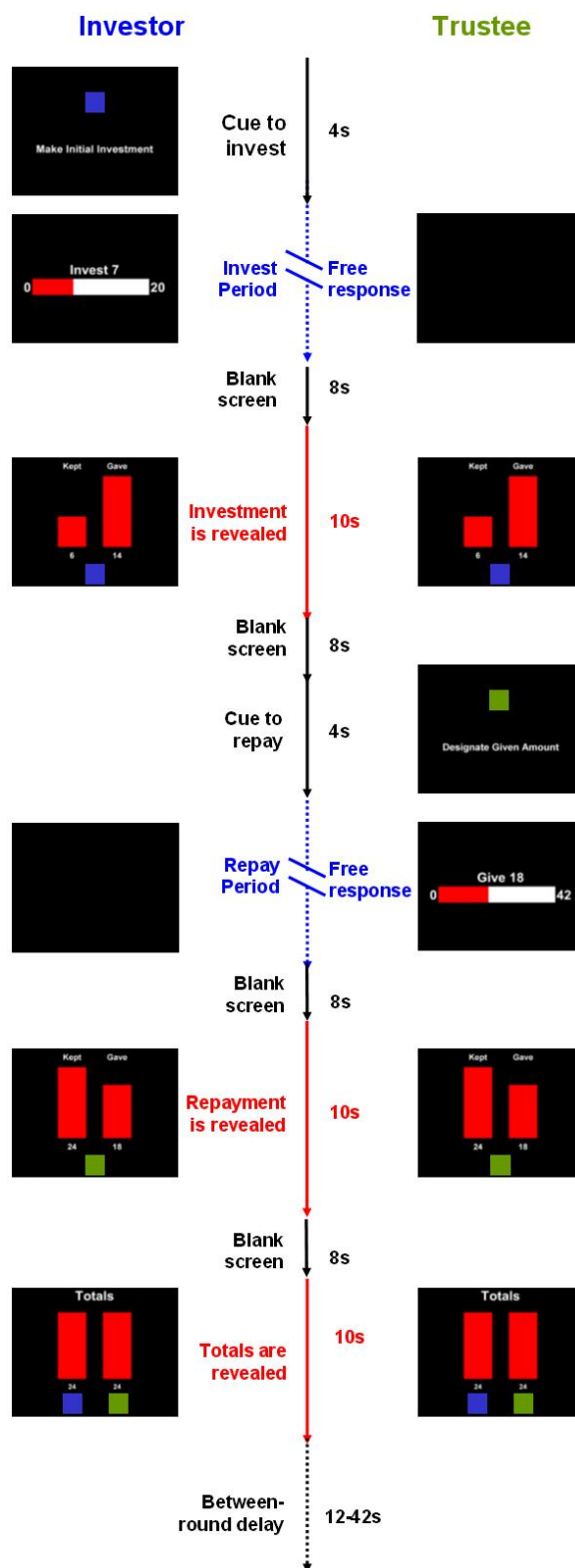


Fig. 20: Timeline of 1 round of the trust game

III.3. Behavioral Results

Since subjects were free to invest/return as much money as they wanted in each round, there was a lot of inter-pair variability, resulting in a rich behavioral space. A few examples of typical money exchanges are presented below.

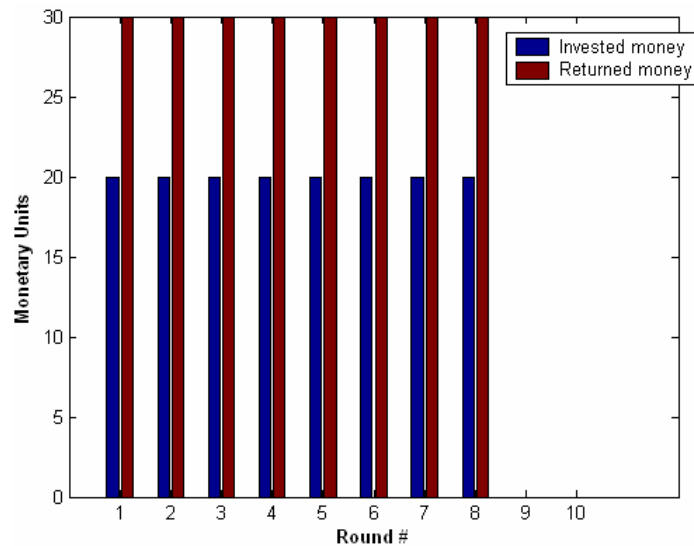


Fig. 21: Example #1 of monetary exchange

Figure 21 shows the interaction between two cooperating subjects. The Investor invests his whole endowment in every round, and the Trustee repays his trust by splitting the money equally between the two. In round 9 however the Investor invests nothing, as he is probably worried that the Trustee might not pay anything back this close to the end of the game. The Investor seems to have done two steps of iterated reasoning with respect to the game-theoretic solution.

Figure 22 shows an initial build-up of trust based on reciprocity: in rounds 1 to 3 both players invest and return increasingly more monetary units. The trust relationship lasts until round 9 when the Trustee suddenly takes 59 out of the 60 available monetary units. This

time it was the Trustee who did the 2 steps of iterated game-theoretic reasoning. In response to the betrayal the Investor invests nothing in Round 10.

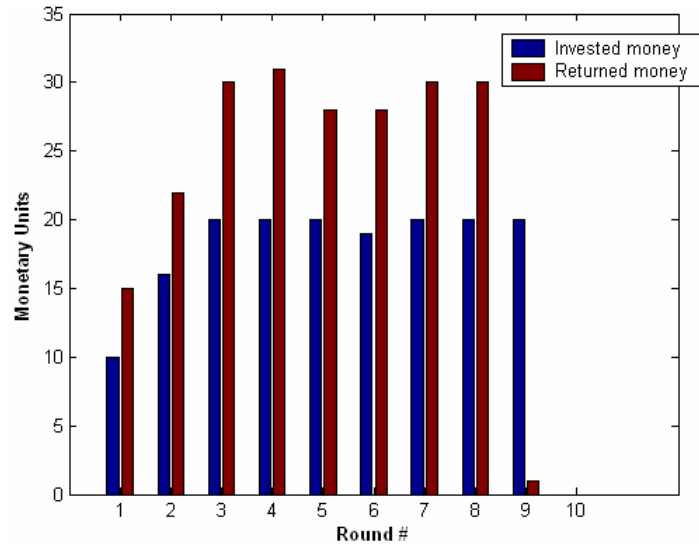


Fig. 22: Example #2 of monetary exchange

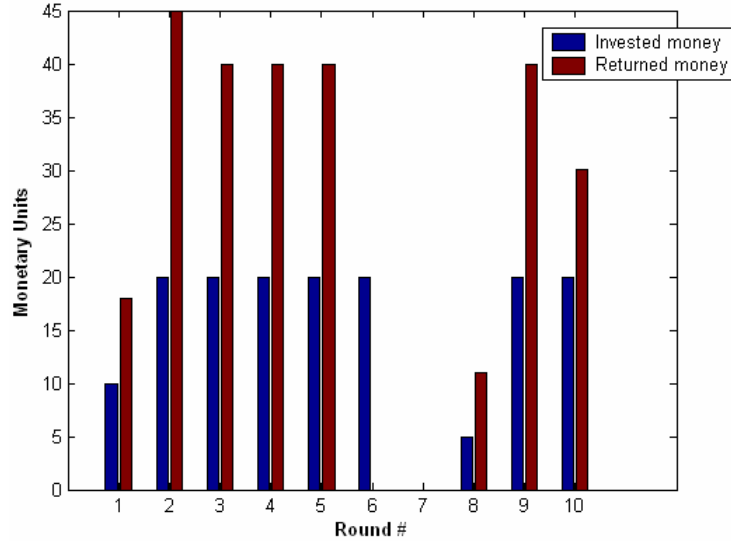


Fig. 23: Example #3 of monetary exchange

In the exchange shown in Figure 23, the players establish a trust relationship quickly, but in round 6 the Trustee defects and takes everything. He is immediately punished as the

Investor invests nothing in Round 7. The Investor seems to be forgiving as he starts investing again in round. Typically players give their opponent a second chance if they split the money in an unfair way. Also note that in this example neither player takes all the money in the last round(s) unlike in the two previous examples.

The average behavior of all 48 subject pairs is shown in Figure 24. Cooperation between players was the strongest during middle rounds: trusting behavior (identified by large investment ratios), peaked during round 6 when Investors invested an average of 81% of the available money, and trustworthiness (identified by large repayment ratios) peaked in round 4 when Trustees returned an average of 47% of the invested money.

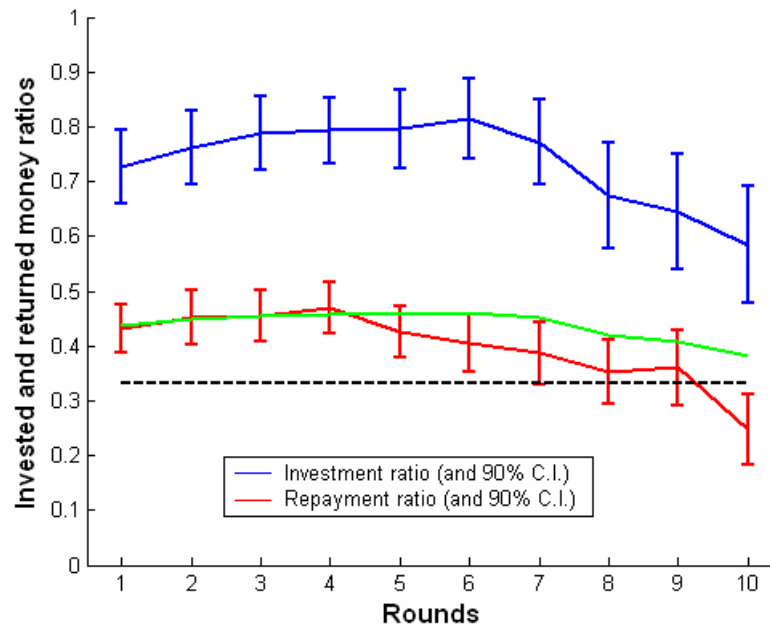


Fig. 24: **Average investment and repayment ratios.** Mean and 90% confidence intervals of the investment ratio (Money sent to Trustee/20) and the repayment ratio (Money sent to Investor / 3*Money sent to Trustee) over the 10 rounds of the trust game. The dotted black line shows the minimum repayment ratio for which the investor is guaranteed to get at least his investment back. The green line indicates how much the repayment ratio needs to be such that the Investor and Trustee both earn the same amount of money. A repayment ratio below the green line indicates that the Trustee earned on average more than the Investor in that round.

After these peaks, both investment and repayment ratios declined over rounds, reflecting a decrease in cooperation. They reached their lowest value in the last round, which was also the only round where investments did not earn a profit on average (the repayment ratio is lower than 1/3). No Investor played the perfectly selfish Nash equilibrium in which the Investor invests nothing in each round. On average Investors earned \$256.54 +/- 56.08 and Trustees earned \$237.58 +/- 63.42, both resulting in an actual payoff of \$35.

III.4. Strategic Uncertainty and Prediction Analysis

III.4.1. Background

Human social life depends on the ability to predict the likely behavior of others, a capacity that underlies cooperation and social institutions (Henrich, Boyd et al. 2005). Unlike individual decision-making under uncertainty, the hallmark of decision-making in social interactions is strategic interdependence. That is, one's best strategy depends on the strategy others adopt, often in response to one's own behavior. This strategic interdependence introduces a novel form of uncertainty, referred to as strategic uncertainty, which is the uncertainty associated with inferring the beliefs and possible actions of others. In most social interactions, we lack perfect information about what others believe, and so lack perfect foresight about how others will respond to our own behavior, creating uncertainty about our predictions. Because of this, strategic uncertainty is a pervasive feature of human strategic interaction—including negotiation, international relations, and trading in such institutions as asset markets—and remains even when all other sources of uncertainty (structural uncertainty) surrounding a decision context are removed (Brandenburger 1996).

To date, little is known regarding the neural basis of strategic uncertainty. The neural basis of social interaction has been explored primarily through investigations of theory of mind (ToM), the capacity to attribute mental states, including beliefs, desires, and intentions to

others. Such work has identified the anterior paracingulate, superior temporal sulci, and temporal poles as regions implicated in ToM (Fletcher, Happe et al. 1995; Goel, Grafman et al. 1995; Baron-Cohen, Ring et al. 1999; Gallagher, Happe et al. 2000). While ToM is thought to have evolved as a capacity to predict the behavior of others through the attribution of mental states that play a role in generating behavior (Premack and Woodruff 1978), ToM can be invoked in situations that do not involve strategic interaction and predictions of future behavior. For example, understanding some forms of humor and retrospectively explaining behavior may require mentalizing abilities but does not involve strategic interaction in the sense that the required mentalizing does not involve predicting a future response to one's own behavior (Gallagher, Happe et al. 2000). For this reason, many ToM studies utilize tasks that require subjects to retrospectively judge a social scenario in which they are not directly involved, thus evoking social, but not strategic, interaction. Thus, while ToM is the capacity to attribute mental states generally, strategic uncertainty is more specifically a form of prediction risk, namely the uncertainty associated with the future response to one's own behavior. Investigation of strategic uncertainty thus requires tasks in which subjects strategically interact with one another, rather than make social judgments retrospectively.

Another potential difference between ToM and strategic uncertainty is that ToM is often regarded as a form of cognitive judgment (a folk theory) with cortical substrates (Fletcher, Happe et al. 1995; Goel, Grafman et al. 1995; Baron-Cohen, Ring et al. 1999; Gallagher, Happe et al. 2000). In contrast, it is possible that strategic uncertainty may be more related to other forms of uncertainty processing, many of which have subcortical substrates (Preuschoff, Bossaerts et al. 2006). This possibility suggests that strategic uncertainty may invoke neural structures that are distinct from those invoked in ToM tasks and may overlap with those involved in uncertainty processing. In recent years, a convergence between reinforcement learning models (Sutton and Barto 1998), primate physiology (Fiorillo, Tobler et al. 2003), and fMRI (Preuschoff, Bossaerts et al. 2006) has made significant progress in understanding how midbrain dopamine structures encode the uncertainty that arises when stimulus-reward association are probabilistic and changing over time. Thus,

whereas reported activations related to ToM have been cortical, this thesis examines whether strategic uncertainty in a high-level social exchange may evoke similar subcortical structures.

A further salient difference between ToM and strategic uncertainty is that strategic uncertainty is quantifiable, whereas ToM typically is not. Since strategic uncertainty is the predictability of a decision-maker's response to one's decision, this predictability can be quantified by the entropy of one's strategic choice, as different strategies often differ in the predictability of the response they will evoke. For example, when a car dealership advertises the purchase price of a car, both very high and very low prices involve relatively little strategic uncertainty, since very high prices have a low probability of acceptance and very low prices have a very high probability of acceptance. If the car dealership is motivated to sell the car for the highest possible price, the task becomes that of predicting the highest price that will be accepted, which will involve relatively high levels of strategic uncertainty. Thus, a goal of this study was to examine whether the entropy of different strategies related in a parametric manner to the magnitude of neural activations, which would provide strong evidence that such activations encoded strategic uncertainty.

Since entropy relates strategic uncertainty to different strategies, it was also interesting to see whether future strategic choices could be accurately predicted simply from the magnitude of neural signals relating to strategic uncertainty. Some studies have investigated whether neural signals were predictive of certain perceptual events (Kamitani and Tong 2005), but to our knowledge no studies to date have examined whether simply knowing the magnitude of neural signals is sufficient to predict future behavioral decisions.

A hallmark of strategic uncertainty is that it is typically maximal when decision-makers have no information about other decision-makers. Hence, successful strategic interaction depends on reducing strategic uncertainty through learning how to predict the behavior of others, and particularly learning how to predict their likely responses to one's own strategies. Such learning has been intensively investigated in theories of learning in game theory (Fudenberg and Levine 1998), which provides a framework to investigate how

players modify their strategic behavior by calibrating strategic uncertainty as they learn from responses of others across repeated plays.

Based on the above considerations, we investigated strategic uncertainty in the trust game. Specifically, we examined whether there are brain signals that both reflect strategic uncertainty as a function of strategic learning in previous rounds and predict future strategic choice.

III.4.2. GLM Analysis: New vs. Known Information

The structure of the finitely repeated trust game requires subjects to modify their strategic behavior as the game evolves and information in the form of opponent responses becomes available. In particular, two critical moments of a round influence strategic choice: the revelation of the results from the investment and repayment phases.

Cluster			Voxel				Area				
P_{cor}	KE	P_{unc}	pFWE	pFDR	T	(Z)	X	Y	Z	L/R	Area
0.042	163	0.023	0.001	0.003	5.99	5.14	3	9	48	R	Sup. Fr. Gyrus, BA6/8
0.099	109	0.057	0.004	0.003	5.63	4.90	9	3	0		Caudate
			0.026	0.004	4.97	4.43	21	9	-6		
0.288	49	0.186	0.004	0.003	5.60	4.87	-42	21	-15	L	Orbitofr. Cortex, BA47
			0.012	0.003	5.26	4.64	-33	18	-15		
0.238	59	0.149	0.008	0.003	5.39	4.73	-9	6	0	L	Caudate
0.387	34	0.267	0.024	0.004	5.00	4.46	-27	-99	0	L	Visual Cortex
			0.189	0.012	4.21	3.85	-33	-93	-6		
			0.687	0.045	3.49	3.27	-15	-99	0		
0.149	85	0.088	0.027	0.004	4.96	4.42	0	36	-12		Med. Fr. Gyrus, BA11/32
			0.517	0.030	3.69	3.44	12	30	-12	R	
			0.676	0.044	3.50	3.28	0	54	-6		
0.494	22	0.373	0.091	0.008	4.50	4.09	36	15	-15	R	Orbitofr. Cortex, BA47
0.599	13	0.499	0.117	0.009	4.40	4.01	45	-84	-3	R	Visual Cortex
0.317	44	0.209	0.194	0.013	4.19	3.85	6	-93	-6	R	Visual cortex
			0.362	0.020	3.90	3.61	24	-99	-6		
0.464	25	0.341	0.203	0.013	4.17	3.83	-3	-27	-3	L	Midbrain

Table 1: **Investment > repayment regions for the Trustee.** Local maxima of clusters in the Trustee brain that show increased activity for the revelation of the investment results relative to the repayment screen ($df=47$, $p<0.001$ uncorrected, cluster size $k\geq 8$). P_{cor} =corrected (family-wise) cluster-level p-value; KE = cluster size (voxels); P_{unc} = uncorrected cluster-level p-value; pFWE = corrected (family-wise) voxel-level p-value; pFDR = corrected (false-discovery rate) voxel-level p-value; T=T-statistic of voxel; (Z)=Z-score of voxel; X, Y, Z = MNI coordinates of voxel location (mm); L/R = laterality (L=Left, R=Right)

Although the revelation screens are visually identical, the displayed information is asymmetric in that for the Trustee the investment screens contains new information (the amount invested by the Investor) whereas the repayment screen displays known information (note that this is exactly the opposite for the Investor). As this new information is the basis for subsequent strategic behavior, we defined two regressors of interest, β_{inv} and β_{rep} , corresponding to the revelation of the investment and repayment screens respectively. A general linear model (GLM) analysis was used to identify brain areas in the Trustee brain whose blood oxygenation level-dependent (BOLD) response was greater for the information bearing screen than for the known screen (see Section III.4.8 for methods).

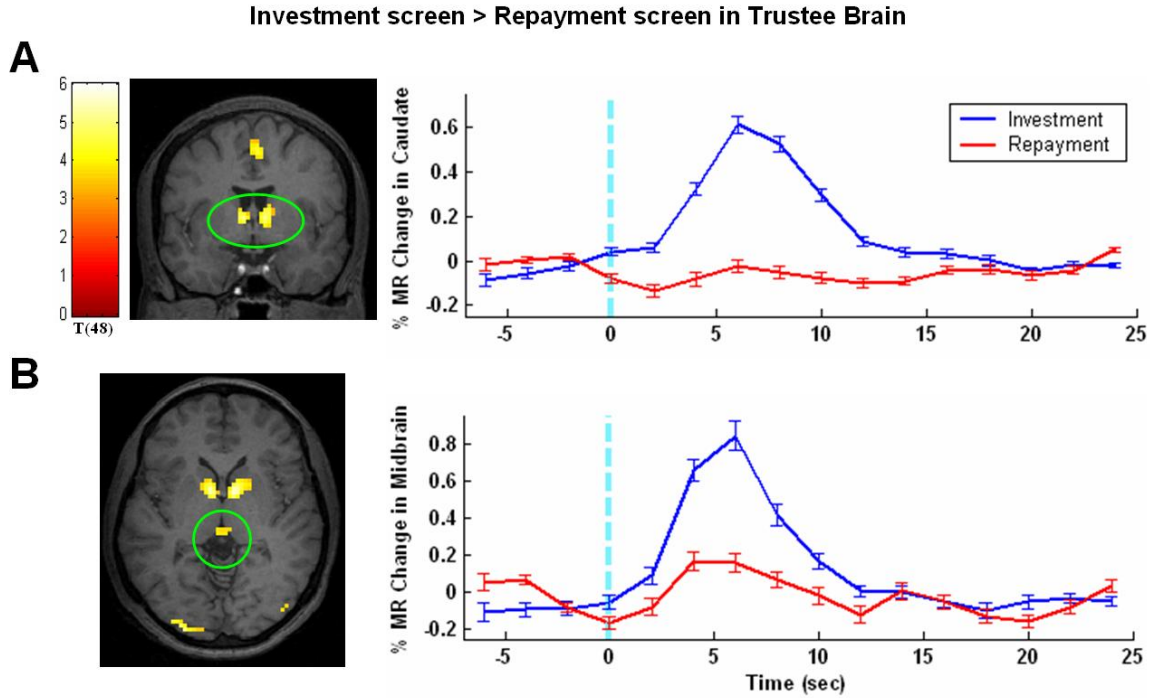


Fig. 25: **Activations in the Trustee Brain.** (A) Left panel: coronal view ($y=0$) of the Trustee brain showing significant differential activation in the bilateral caudate for the contrast $\beta_{inv} - \beta_{rep}$. A statistical map is shown alongside a pseudo-color legend with t-scores ($p \leq 0.001$, minimum cluster size: 8). Right panel: Average time-courses at the moment of the revelation ($t=0$) of the investment and repayment screens in caudate. (B) Similar as in (A), the left panel shows an axial ($z=0$) view of the Trustee with activation in the medial midbrain, and the right panel shows the corresponding average time-courses.

Five regions, all previously implicated in reward-processing and decision-making (Schultz, Dayan et al. 1997; Cohen, Botvinick et al. 2000; Elliott, Friston et al. 2000; Paulus, Hozack et al. 2002; Delgado, Miller et al. 2005), showed significant activation and were used as regions of interest (ROI) for subsequent analysis: bilateral caudate; medial midbrain; superior frontal gyrus (BA6/8), bilateral orbitofrontal cortex (BA47) and medial frontal gyrus (BA11/32) (Table 1 and Figure 25).

III.4.3. ROI Analysis: Correlation between Later Decisions and Hrf

We next investigated whether activation in these ROIs was correlated with subsequent Trustee responses to Investor behavior (see Section III.4.8. for methods on ROI analysis). Consequently, we examined future changes in repayment ratio ΔR_i as a function of current changes in investment ratio ΔI_i (Fig. 26).

When the investment is revealed at round i to Trustee, the current change in investment ratio is defined as $\Delta I_i = I_i / 20 - I_{i-1} / 20$, where I_i and R_i are the investment and repayment amounts respectively at round i . Similarly, the future change in repayment ratio is defined as $\Delta R_i = R_i / (3I_i) - R_{i-1} / (3I_{i-1})$. This representation segregated the behavioral space into four quadrants, each reflecting a strategic response the Trustee may make as a function of Investor behavior (Fig. 26). From a game-theoretic perspective, reciprocal events (green quadrants) reflect tit-for-tat strategies, which are a robust way to create human cooperation in repeated games (Axelrod and Hamilton 1981). Non-reciprocal events (red quadrants) are the result of altruistic (benevolent) and greedy (malevolent) strategies. Neutral events (blue circle) occur when both players have reached some stable pattern of exchange.

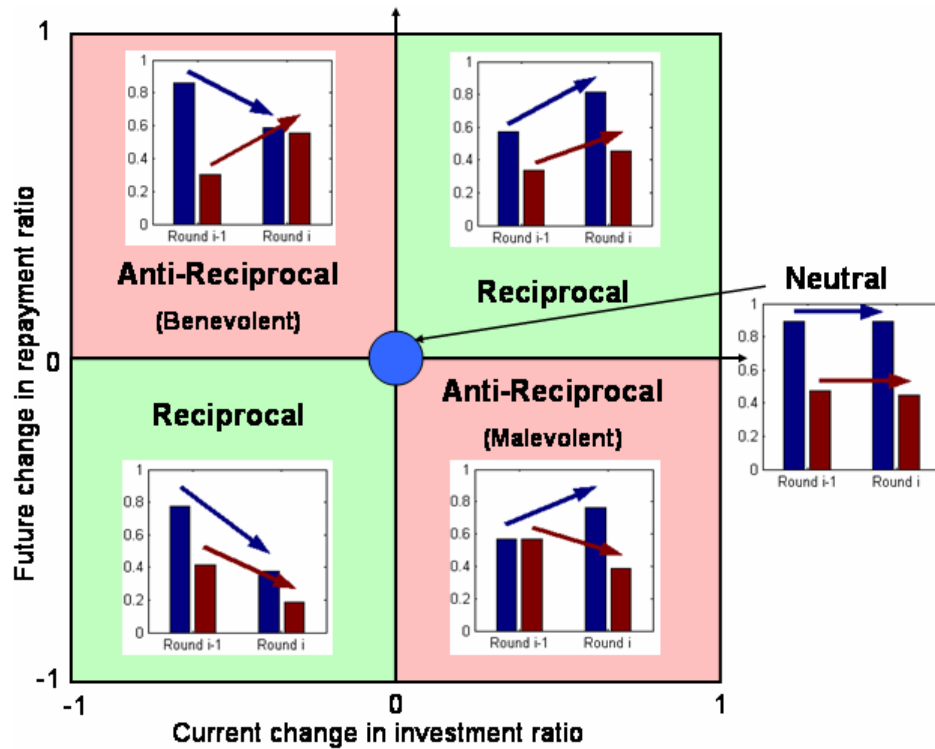


Fig. 26: **Split-up of the behavioral space.** Green quadrants contain reciprocal events and red quadrants contain anti-reciprocal events. Neutral events are located at (0,0) in the blue disk. Subpanels show the average Investment (blue) and Repayment (red) ratios in each quadrant for rounds $i-1$ and i . The blue and red arrows show the direction of the corresponding changes in Investment and Repayment ratios respectively (e.g., in the top-left quadrant the Investor increases his investment ratio, and the Trustee subsequently decreases his repayment ratio).

We next segregated hemodynamic responses in ROIs according to these 3 categories. As we were interested in how brain responses might change as a function of strategic learning during gameplay, we examined these signals across middle (3–6) and late (7–10) rounds. The rationale for this division was based on the hypothesis that by middle rounds players had acquired sufficient information regarding other players' likely responses as a basis for strategic choice and was also supported by evidence from work presented in Section III.6. This analysis revealed that among previously identified ROIs reciprocal, non-reciprocal, and neutral strategies could be distinguished by time-courses in bilateral caudate and midbrain in both middle and late rounds (left and middle panels in Fig. 27). Although the scanning protocol was not optimized for midbrain (Guimaraes, Melcher et al. 1998), the

similarity between caudate and midbrain time-courses led me to conclude that the midbrain activation was reliable.

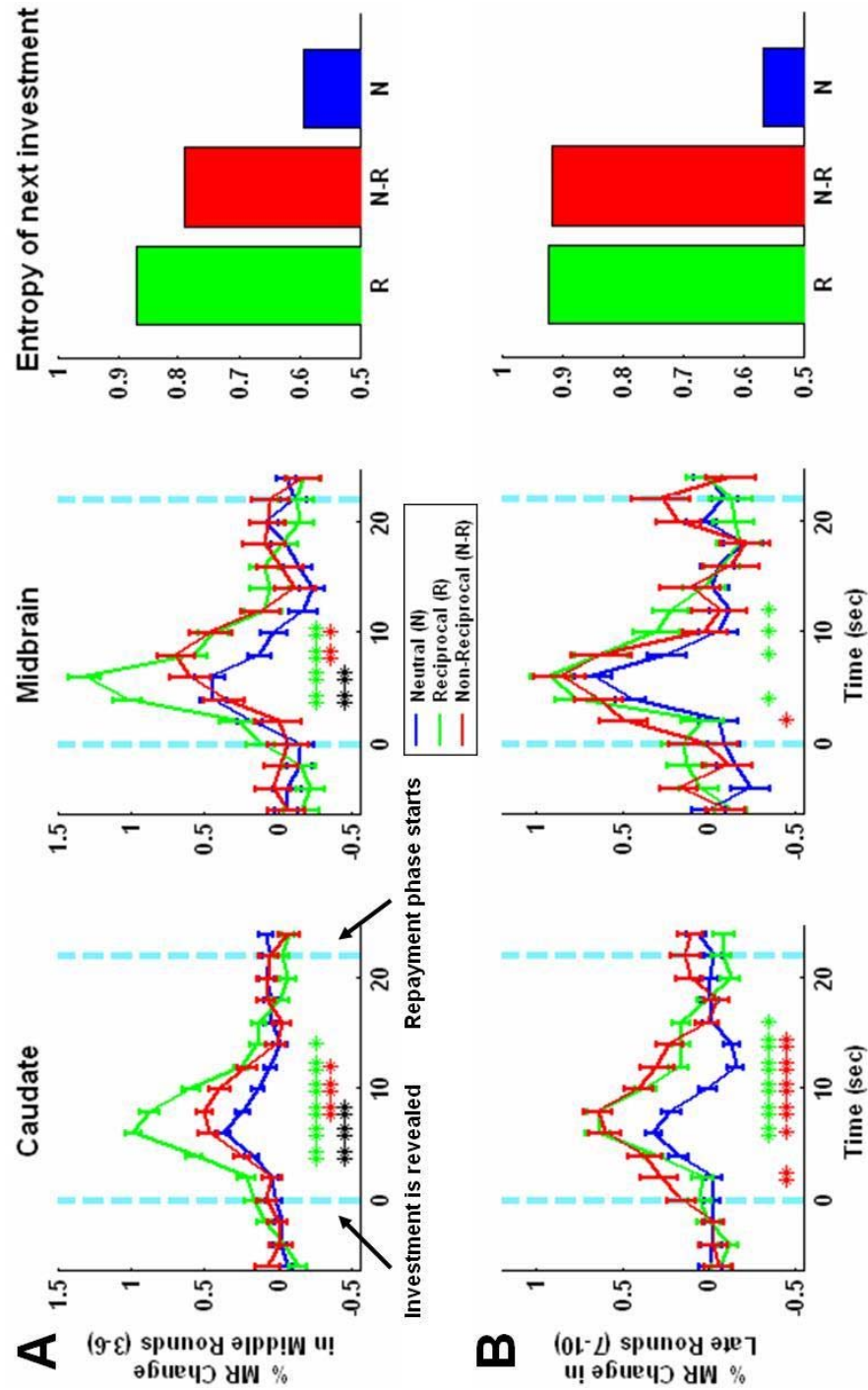


Fig. 27: Time-courses predict strategic choice and correlate with strategic uncertainty. (A) Left & middle panels: time-courses

in bilateral caudate and midbrain during *middle* (3–6) rounds. We segregated signals in response to the revelation of the investment screen ($t=0$) with respect to reciprocal (green), non-reciprocal (red), and neutral (blue) strategies (significance levels: $p<0.05$: *, $p<0.005$: **, R vs. N: green stars, N-R vs. N: red stars, N-R vs. R: black stars). Since the actual repayment phase only starts 22 seconds later, the time-courses are predictive of the trustee's strategy. Right panel: entropy of the Investor's next move as a function of strategy ($H_{\text{reciprocal}}$, $H_{\text{non-reciprocal}}$, H_{neutral}). High entropy values denote high uncertainty levels. These data show that during middle rounds $H_{\text{reciprocal}} > H_{\text{non-reciprocal}} > H_{\text{neutral}}$, which is exactly the same order as signal magnitudes in the left and middle panels. **(B)** Similar as in (A), we segregated signals in the caudate (left panel) and midbrain (middle panel) according to strategy in *late* (7–10) rounds. During late rounds it is no longer possible to distinguish between reciprocal and non-reciprocal strategies. This trend is also repeated in the uncertainty of the Investor's next move, where $H_{\text{non-reciprocal}}$ has increased to the level of $H_{\text{reciprocal}}$. This change in both signal magnitude and entropy from middle to late rounds suggests that signal magnitude in the caudate and midbrain encodes for uncertainty about the Investor's future moves.

During middle rounds time-courses for all three strategies had different peaks: reciprocal events had the largest magnitudes and neutral events had the lowest. By late rounds the amplitude of non-reciprocal events had risen to the level of reciprocal events. This difference between signals across late and middle rounds confirmed the hypothesis regarding strategic learning and the opportunity for players to learn about their opponents in contexts that encouraged reciprocity (Fehr and Gächter 1998) (repeated play of the middle rounds), or discouraged reciprocity (the shadow of the game's end, a phenomenon known as endgame effects, which are well-established empirically in behavioral game theory (Camerer and Weigelt 1988; Dal Bo 2005).

III.4.4. Signal Magnitude Encodes Strategic Uncertainty

We next investigated the relation between the signal magnitudes in bilateral caudate and midbrain and strategic choice by examining why the magnitude of reciprocal strategies was the largest, and why the magnitude of non-reciprocal events increased from middle to late rounds. Our first hypothesis was that signal magnitude may encode expected future reward. However, we found no statistically significant correlation between signal magnitude in the Trustee brain at round i and money received by the Trustee at round i or $i+1$. We thus

tested the alternative hypothesis that signal magnitude encodes the uncertainty of future reward, or strategic uncertainty, i.e., the predictability of the next investment after the Trustee's repayment decision, which can be measured by entropy (Shannon 1948) (see Section III.4.8. for methods). We found a correspondence between the relative magnitudes of brain signals in midbrain and caudate and the entropy of different strategies in middle rounds (rightmost panels in Fig. 27). Both signal magnitude and entropy are the highest for reciprocal events, and the lowest for neutral events. During late rounds the entropy of non-reciprocal events increased to the level of reciprocal events, as did the signal magnitude in midbrain and bilateral caudate (left and middle panels in Fig. 27). This relationship provided further support that the late rounds induce different strategic interactions due to increasing strategic uncertainty as the end of the game draws near. Although players did not fully follow backward induction as predicted by analytical game theory (Camerer 2003), by late rounds they anticipated the final round, reducing the incentive for reciprocity. The correspondence between entropy and signal magnitudes across events in the middle periods, and the increase in both entropy and signal for non-reciprocal events from middle to late periods, strongly suggest that these brain areas encode Trustee strategic uncertainty as measured by entropy.

III.4.5. Signal Magnitude Predicts Future Strategic Choice

Given the correlation between signal magnitude in caudate and midbrain and future strategic choice that occurs substantially later, it suggested the possibility that these signal magnitudes were predictive of strategic choice. We thus examined how accurately future strategic decisions could be predicted on a trial-by-trial basis based on brain activation alone. At time-points when reciprocal and non-reciprocal signals were statistically different, their signal magnitudes followed two distinct normal distributions (Fig. 28A), which led us to hypothesize that these two types of events could be effectively separated. We thus calculated how the probability of non-reciprocal events changed as a function of upperbound signal magnitude, and found that it increased from 34% to 70% as the upperbound signal magnitude decreased from 2 to 0 in the left caudate (Fig. 28B). Since

the behavioral data revealed that the percentage of non-reciprocal events is 34% (independent of signal magnitude), this confirmed our hypothesis that low signal magnitudes are much more indicative of non-reciprocal events.

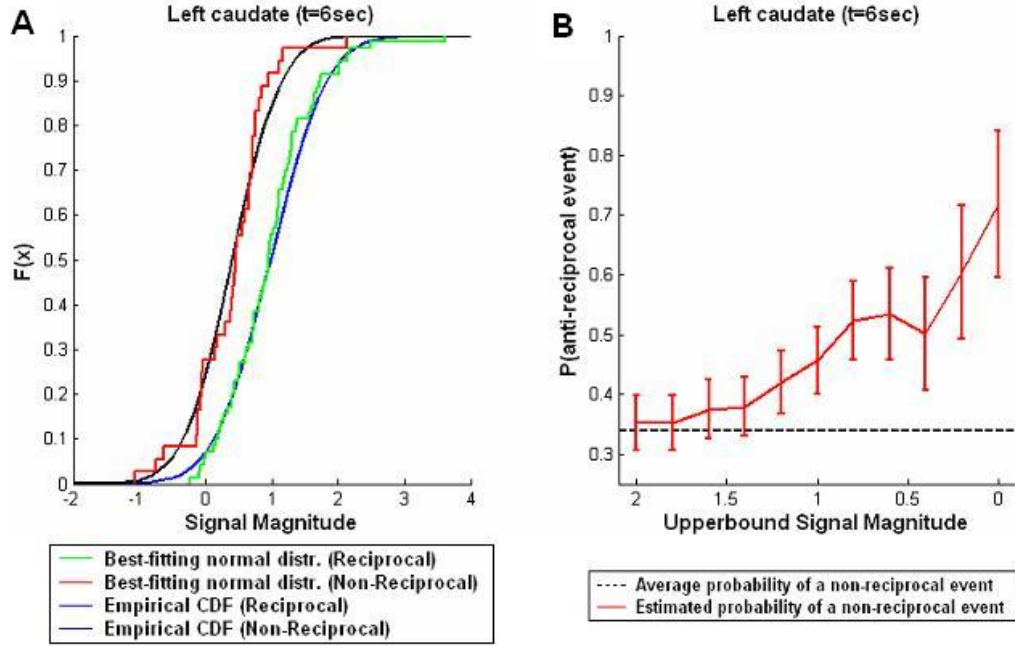


Fig. 28: **Analysis of Signal magnitudes.** (A) Distribution of Signal Magnitude. Cumulative distribution function (CDF) of the signal magnitudes in the left caudate for reciprocal and non-reciprocal events 6 seconds after investments were revealed. In addition to following normal distributions (the Lilliefors test for goodness-of-fit to a normal distribution with unknown mean and unknown variance could not be rejected at the 5% level), the distributions of the peaks are also statistically different. (Kolomogorov-Smirnov test: left caudate $p=0.000024$; right caudate $p=0.006$; midbrain $p=0.0004$, all for $t=6$ sec). (B) Probability of a Non-reciprocal Event in the Left Caudate. We calculated the probability of a non-reciprocal event as a function of upperbound signal magnitude t : $P_t = Prob(x \text{ is an non-reciprocal event and } s_x \leq t)$ which can be estimated by:

$$P_t \approx \frac{\#(NR \leq t)}{\#(NR \leq t) + \#(R \leq t)} \text{ where } s_x \text{ is the signal magnitude}$$

of event x , and $\#(NR \leq t)$ and $\#(R \leq t)$ represent the number of non-reciprocal respectively reciprocal events whose signal magnitude is less than t . Error bars were obtained through bootstrap, by sampling the data with replacement ($N=1000$). As the signal upperbound magnitude decreases, this probability increases and is significantly above the average probability of a non-reciprocal event (black dotted line). Results in the right caudate and in the midbrain look very similar, and are therefore not displayed here.

III.4.6. *Discriminant Analysis of Predictive Accuracy*

We were also interested in designing a method to robustly predict reciprocal and non-reciprocal events based on signal magnitudes alone. From the many possible data classification techniques that exist (e.g., support vector machines, principal component analysis), we implemented Fisher Linear Discriminant Analysis (F-LDA) (Fisher 1936; Mika, Rätsch et al. 1999; Seber 2004) as it provides feature extraction for classifying data rather than feature extraction for describing data (as does principal component analysis). F-LDA finds the feature that best discriminates between 2 classes (here: reciprocal and non-reciprocal events) by maximizing the between class variance while minimizing the within class variance (see methods).

To increase the prediction accuracy despite the high BOLD variance, we included signal magnitudes from up to 9 different time-points in midbrain and bilateral caudate, and used all reciprocal and non-reciprocal events from middle rounds. We used the jack-knife method (Quenouille 1956) to cross-validate the results by successively applying F-LDA to all but one event. To verify whether the left-out event was correctly classified, we calculated a classification score based on the Mahalanobis distance in the new feature space obtained from the F-LDA (see Section III.4.8. for methods).

As more time-points were included in the analysis, the prediction performance improved from 72.64% for 2 time-points (Fig. 29A) up to 78.30% for 6 time-points (Fig. 29B) from both caudate and midbrain. This percentage is significantly above the chance level of 50%, when no priors are taken into account (i.e. no assumption about the ratio of reciprocal to non-reciprocal events is being made). If priors are taken into account, the chance level increases to 66% (the proportion of reciprocal events), which is still considerably below the prediction performance of 78.3%.

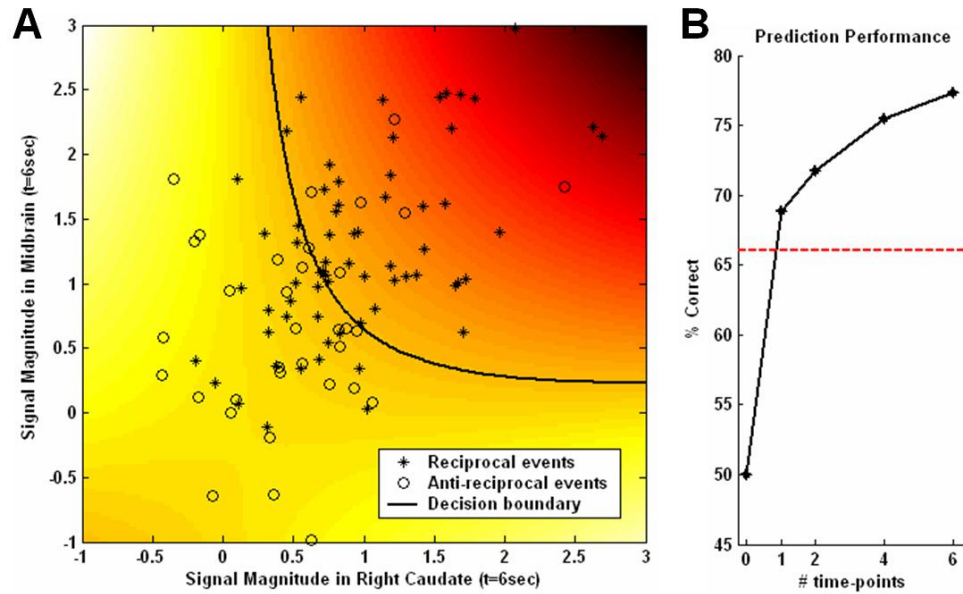


Fig. 29: **Performance of the Prediction Analysis.** (A) Two-dimensional decision space based on a Fisher linear discriminant analysis (F-LDA). F-LDA was used to transform signal magnitudes in the right caudate and midbrain into a new feature space that discriminates maximally between reciprocal and non-reciprocal events. The new space was split up into reciprocal and non-reciprocal partitions using the Mahalanobis distance (see Section III.4.8 for methods) and transformed back into the signal magnitude space (plotted here). The resulting decision boundary splits the signal magnitude space into a reciprocal partition (reddish colors) and a non-reciprocal partition (yellowish colours). Actual reciprocal and non-reciprocal events are encoded by asterisks and circles respectively. (B) Performance of the discriminant classification algorithm as dimensionality increases. The prediction performance is considerably above the chance level of 66% (red dotted line) which takes into account the prior distributions of reciprocal and non-reciprocal events.

III.4.7. Discussion and Conclusion

We used an anonymous social exchange task and event-related fMRI of interacting subjects to investigate the neural correlates of strategic interaction. The results demonstrate that brain signals in the caudate and midbrain encode strategic uncertainty that reflects both strategic learning and game context and predicts future strategic choice in a repeated game. In relating the signal magnitude in caudate and midbrain to strategic uncertainty, it may seem counterintuitive that non-reciprocal strategies (benevolence and malevolence) lead to

less uncertainty about the investor's next move than tit-for-tat strategies. However, despite combining events of opposite valence, non-reciprocal strategies send a very strong signal to the Investor: benevolent strategies lead the Investor to increase investments, and malevolent strategies lead Investors to decrease investments, both leading to reduced uncertainty about the Investor's next move.

A highly salient event in the repeated trust game for the Trustee is the revelation of the investment screen, as this screen reveals critical information about the Investor, the Investor's response to the Trustee's previous repayments, and provides a basis for subsequent strategic choice. The GLM analysis revealed a complex response to this screen in the Trustee brain, including cortical activations in BA6/8 and orbitofrontal cortex (BA47). Despite being regarded as a traditional "motor" area, previous studies have reported BA6/8 activation during non-motor cognitive tasks (Paulus, Hozack et al. 2002; Tanaka, Honda et al. 2005), and elicits increased activation for predictions under uncertainty (Volz, Schubotz et al. 2003). The orbitofrontal cortex (BA47) has been reported to mediate emotional influences on decision-making (Damasio 1994) and to adapt responses to different behavioral contingencies (Rolls 1996). It has also been activated in risk-taking processes under different psychological contexts (Cohen, Botvinick et al. 2000). BA11/32 has been linked to social decision-making (Elliott, Friston et al. 2000; Rilling, Gutman et al. 2002), in particular during cooperative behavior (McCabe, Houser et al. 2001).

Therefore, the cortical activations we report in the Trustee brain to this salient event are in broad agreement with previous reports linking these cortical areas to uncertainty and social interaction. However, it is intriguing to note that these cortical activations were not correlated with future Trustee responses. Only signal magnitudes in caudate and midbrain, structures previously implicated in reward-processing in individual decision-making, predicted what type of strategy (reciprocal, non-reciprocal, or neutral) the Trustee would pursue later in the round. It is remarkable that strategic choice in a high-level social exchange can be predicted from a signal in subcortical structures, whereas reversal learning

and other forms of cognitive control have been attributed to prefrontal structures (Remijne, Nielen et al. 2005). In Section III.6 the involvement of the caudate in trust is reported (King-Casas, Tomlin et al. 2005), which is also traditionally regarded to be a high-level process. The involvement of subcortical structures in strategic choice, however, may be due to the involvement of these structures in prediction learning. The midbrain dopaminergic system plays an important role in the brain reward system (Schultz, Dayan et al. 1997) and has been shown to make predictions about likely rewards. Recent findings have identified reward prediction error signals from reinforcement learning in the human caudate and putamen (McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003; Seymour, O'Doherty et al. 2004; Knutson and Cooper 2005; Haruno and Kawato 2006) that are thought to involve outputs of the midbrain dopaminergic systems. It may, therefore, be another striking feature of these subcortical structures, which are highly conserved across species (Montague, Dayan et al. 1995), that they have been recruited to subserve novel forms of prediction learning in human social exchange. It may also be the case that these structures primarily mediate learning during strategic interaction, and that once a predictive model of a strategic partner is built, other structures are recruited (Delgado, Frank et al. 2005). This learning phase might be crucial for strategic interaction, and crucial for updating models of strategic partners when predictions fail, consistent with the notion of prediction error-driven learning (Sutton and Barto 1998). Evidence from electrophysiological and fMRI data support our hypothesis that the midbrain encodes uncertainty. Recordings from dopaminergic neurons in the monkey midbrain have been shown to correlate with reward uncertainty (Fiorillo, Tobler et al. 2003; Volz, Schubotz et al. 2003), and activation in the midbrain has been correlated with the entropy of the outcomes during a classification learning task (Aron, Shohamy et al. 2004). However, in those cases the uncertainty was always associated with an action or stimulus that had already been chosen or presented before—here we present the novel element that midbrain activity is correlated with uncertainty about a future action. Support for the predictive nature of the midbrain activity comes from a recent electrophysiology study that showed that the activity of midbrain dopamine neurons reflects future choice as early as 122 ms after the presentation of a conditioning stimulus (Morris, Nevet et al. 2006). However, they

do not correlate the predictive activation with uncertainty, but favor the hypothesis that the dopamine neurons receive information about the decision from another structure. Although we cannot rule out that possibility, none of the cortical structures that was activated during the reveal screens showed the same predictive activation pattern as the midbrain or caudate.

As subjects build up their reputation during middle rounds, deciding between following a reciprocal or non-reciprocal strategy is a crucial decision. It is therefore striking that strategic choice can be accurately predicted by an encoding of strategic uncertainty in the signal magnitude of caudate and midbrain. More specifically, this signal peaked approximately 16 seconds before the Trustee's actual repayment decision began and approximately 31 seconds before the Trustee's decisions were lodged, the behavioral manifestation determining whether the event is reciprocal or not. During this period, Trustees may be deliberating about their possible response; however, from the point of view of this predictive signal, their subsequent strategic choice has already been determined with high accuracy almost immediately following the revelation of the Investment screen. Previous studies have investigated how fMRI signals in the early visual areas can reliably predict attended orientation (Kamitani and Tong 2005), or how activation in different brain areas precedes risk-seeking and risk-averse behavior (Kuhnen and Knutson 2005); here we have shown how time-courses in reward-related areas strongly predict strategic behavior during a social exchange task, a ubiquitous and critical capacity for strategic interaction in everyday life.

While the neuronal basis of social cognition has previously focused on cortical structures, we have designed an interactive task in which subjects' performance (in the form of real monetary rewards) depends on predicting how others will respond to their own behavior, and have shown that a crucial element of social cognition, strategic uncertainty, is mediated by subcortical structures, midbrain, and caudate. Previously identified with reward learning under uncertainty, the finding that these structures are involved in a high-level social decision-making process and that activity in those structures evolves during game-play,

suggests that the brain may utilize common structures for both non-social and social forms of uncertainty learning. Further evidence for this possibility stems from our finding that the fMRI signal magnitudes in these structures corresponds to strategic uncertainty as measured by entropy and, by using discriminant analysis, accurately predicts a player's future strategic choice. Altogether, these results suggest that the human brain's remarkable social cognition capacities may depend in part on extending computational processes originally used for non-social uncertainty learning to the social domain.

III.4.8. Methods

General Linear Model (GLM) Analysis. GLM analysis (Friston, Holmes et al. 1995) was done in SPM2 by specifying a separate general linear model for each subject (fixed effects analysis). First all images were high-pass filtered in the temporal domain (filter width 128 s) and autocorrelation of the hemodynamic responses was modelled as an AR(1) process. In the GLM model all visual stimuli and motor responses were modeled as separate regressors that were constructed by convolving a hemodynamic response function (hrf) with a comb of Dirac functions at the onset of each visual stimulus or motor response. Following the GLM analysis, a voxel-by-voxel contrast analysis was performed to identify voxels for which $\beta_{\text{inv}} > \beta_{\text{rep}}$, i.e., areas in the Trustee brain that were significantly more activated for the revelation of the investment screen than for the revelation of the repayment screen. Next, a random effects analysis was done on the contrast images $\beta_{\text{inv}} > \beta_{\text{rep}}$ of all 48 trustees by implementing a one-sample t-test. A similar analysis ($\beta_{\text{rep}} > \beta_{\text{inv}}$) revealed that activations in the Investor brain are similar to those found in the Trustee brain (Fig. 30 and Table 2), but a detailed analysis of fMRI time-courses during the repayment screen did not allow us to distinguish between strategic choice, i.e., reciprocal, non-reciprocal, and neutral strategies. We hypothesize that this is due to the long waiting time (~ 1 minute) between the repayment screen and the start of the next investment phase for the Investor.

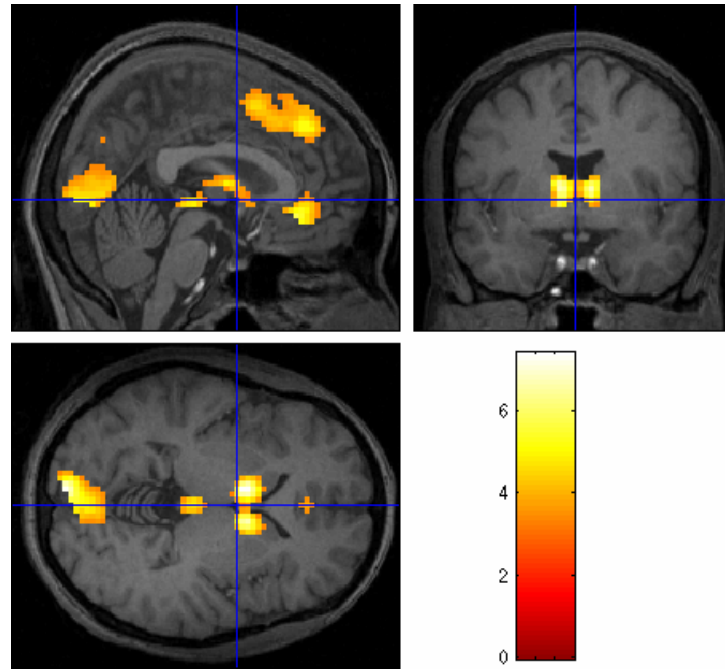


Fig. 30: **Activations in the Investor brain.** Regions in the Investor brain that showed higher activation for the repayment screen than for the investment screen ($x=0$, $y=0$, $z=0$; neurological convention; statistical map shown alongside pseudo-color legend with t-scores, $p \leq 0.001$, $df=46$)

Cluster			Voxel				Area				
P_{cor}	k_E	P_{unc}	P_{FWE}	P_{FDR}	T	(Z)	X	Y	Z	L/R	Area
0.012	330	0.010	0.000	0.000	7.41	5.98	-9	6	0	L	Caudate
			0.000	0.000	6.68	5.56	9	3	3	R	
0.010	356	0.008	0.000	0.000	7.19	5.86	-9	-96	0	L	Visual Cortex
			0.020	0.001	4.92	4.39	-6	-84	6		
			0.060	0.003	4.51	4.08	0	-75	3		
0.164	88	0.149	0.002	0.000	5.66	4.90	0	33	-9		Med. Fr. Gyrus, BA32
0.258	55	0.248	0.006	0.001	5.36	4.70	-3	-24	-3	L	Midbrain
0.055	179	0.047	0.011	0.001	5.13	4.54	0	36	39		Sup. Fr. Gyrus, BA6/8/9
			0.053	0.003	4.56	4.12	0	12	51		Sup. Fr. Gyrus, BA6/8
			0.142	0.007	4.17	3.82	0	24	42		
0.488	15	0.557	0.020	0.001	4.93	4.40	-48	18	-9	L	Inf. Fr. Gyrus
0.548	9	0.659	0.342	0.014	3.76	3.49	-15	-60	6	L	Posterior Cingulate
0.526	11	0.621	0.564	0.026	3.45	3.24	-3	-75	33	L	Parietal Lobe, BA7
			0.593	0.028	3.42	3.21	-6	-78	42		

Table 2: **Repayment > investment regions for the Investor.** Local maxima of clusters in the Investor brain that show increased activity for the revelation of the repayment results relative to the investment screen ($df=46^1$, $p < 0.001$ uncorrected, cluster size $k \geq 8$). P_{cor} = corrected (family-wise) cluster-level p-value; k_E = cluster size (voxels); P_{unc} = uncorrected cluster-level p-value; p_{FWE} = corrected

¹ One subject was excluded because the functional data could not be correctly coregistered to the anatomical scan.

(family-wise) voxel-level p-value; pFDR = corrected (false-discovery rate) voxel-level p-value; T=T-statistic of voxel; (Z)=Z-score of voxel; X, Y, Z = MNI coordinates of voxel location (mm); L/R = laterality (L=Left, R=Right)

Region of Interest (ROI) Analysis. ROI analysis was performed in the midbrain and in the bilateral caudate on the 25 most activated voxels identified in the random effects analysis (except for the midbrain where only 21 voxels passed the desired voxel threshold). For each ROI the time-series of each voxel were extracted from the preprocessed images and the effects of no interest (i.e., all motion responses and visual stimuli other than the display of the investment and repayment screens) were removed using the parameter estimates from the GLM. After synchronizing the time-series to the behavioral times for each round, the percent signal increase was calculated. Next the time-series were averaged spatially (over all voxels in the ROI), temporally (over all rounds within a subject), and finally over all subjects. Hence error-bars on the time-courses were based on a single observation per subject.

Entropy. For each strategic choice (reciprocal, non-reciprocal, and neutral) at round i we determined the probability of an investment increase p_{increase} (when $I_{i+1} - I_i > 1$), decrease p_{decrease} ($I_{i+1} - I_i < -1$), or no change p_{neutral} ($|I_{i+1} - I_i| \leq 1$) in the next round. Using that probability distribution, the normalized entropy of the investor's next move was calculated as $H = -\sum_x p_x \log_3(p_x)$ for each strategy and for middle and late rounds.

Fisher Linear Discriminant Analysis (F-LDA) and Classification. F-LDA is used to transform the data into a discriminant space where reciprocal and non-reciprocal events can be more easily separated. Let $E_R = \{x_j^R\}_{j=1 \dots R}$ and $E_N = \{x_j^N\}_{j=1 \dots N}$ be the sets of reciprocal and non-reciprocal events respectively, where x_j^R and x_j^N are L-dimensional vectors composed of the magnitudes from up to 9 different time-points in bilateral caudate and midbrain (at times $t = 4, 6$, and 8 seconds after the display of the Investment screen). Fisher's linear discriminant is given by the vector v which maximizes:

$$J(v) = \frac{v^t S_B v}{v^t S_W v}$$

where

$$S_B = (\mu_R - \mu_N)(\mu_R - \mu_N)^t \text{ and}$$

$$S_W = \sum_{i=R,N} \sum_j (x_j^i - \mu_i)(x_j^i - \mu_i)^t$$

are the between and within class scatter matrices respectively, and $\mu_i = \frac{1}{i} \sum_j x_j^i$ for $i = R, N$. Maximizing $J(v)$ can be transformed into the following constrained optimization problem:

$$\min_v -v^t S_B v$$

$$\text{s.t. } v^t S_W v = 1$$

which corresponds to the Langrangian:

$$L = -v^t S_B v + \lambda(v^t S_W v - 1).$$

After solving the Langrangian and some matrix manipulations, the problem reduces to a regular eigenvalue problem. The objective $J(v)$ is maximized when the eigenvector \tilde{v} corresponding to the largest eigenvalue is chosen.

In order to estimate the performance of the F-LDA, the leave-one-out jack-knife method was used to classify test events in the new discriminant space. As the time-points are correlated, the Mahalanobis distance was used to measure the distance between the test event and the reciprocal/non-reciprocal group centroids in the new discriminant space. The

test event was classified as belonging to the group to which it is closest, that is, where the Mahalanobis distance is smallest. The Mahalanobis distance between a group of events with centroid $c = (c_1, c_2, \dots, c_L)$ and covariance matrix Σ and a vector $x = (x_1, x_2, \dots, x_L)$ is given by:

$$d_M = \sqrt{(x - c)^T \Sigma^{-1} (x - c)} .$$

The F-LDA and subsequent Mahalanobis-based classification algorithm was applied to signal magnitudes in bilateral and midbrain at the time of the peak of the hrf (at times $t = 4, 6, \text{ and } 8$).

III.5. Dynamic Cross-Brain Analysis

III.5.1. Background

The most commonly used technique to analyze functional magnetic resonance imaging (fMRI) data is the general linear model (GLM), developed by Friston et al. (Friston, Holmes et al. 1995). The GLM analysis is a linear regression method that first minimizes the squared distance between a model specified by the experimenter and an fMRI time-series from a 3D volume element (voxel) in the brain, and then assesses the validity of the model using different statistical tests. Although the GLM has proven to be very successful for detecting activations in the brain, its relatively simplistic nature restricts it from exploring the complexity and richness of fMRI data. Another drawback of the GLM is that it cannot be used in its present form to study the brains of interacting subjects. As neuroscientists are becoming more knowledgeable about how an individual brain functions, an important challenge is to understand the neural correlates of (social) interaction: how do the brains of two or more people function when they interact, cooperate, or deceive each other? In this section we propose a new data analysis method that can be used in addition to

the GLM model to extract further information from the fMRI time-series of two interacting subjects.

Our method makes use of dependency measures to tie strategic behavior exhibited by the players to temporal interactions between fMRI time-series. The continuous exchange of money between the two players in the trust game creates a rich and dynamic strategic behavior. More specifically, we were interested in relating strategic choice to temporal interactions in specific brain areas both within and in-between players. In the following we label this combination of neural and behavioral interactions as game dynamics.

To perform the dependency analysis we used two common tools for assessing linear and non-linear correlations: the correlation coefficient and the mutual information. The idea of using mutual information as a tool for fMRI data analysis is not new. It has been successfully applied as an alternative to the GLM model by making fewer assumptions about the relationship between the model and the fMRI time-series (Tsai, J.W. et al. 1999). It has also been used in functional connectivity studies (Friston 2003; David, Cosmelli et al. 2004) to assess correlations between different structures in the brain. Here I use it in addition to the correlation coefficient to detect non-linear temporal dependencies within a brain area.

In order to explore game dynamics, we first isolated areas of interest using the GLM model, and then applied the dependency analysis to the fMRI time-courses to detect strategic interactions. In the following we start off by describing the theory behind the dependency analysis, then we present the methods and results, and we conclude with a discussion of the main findings.

III.5.2. Dependency Measures

The goal is to quantify the relationship between 2 discrete fMRI time-series $X = (x_1, x_2, \dots, x_N)$ and $Y = (y_1, y_2, \dots, y_N)$ acquired simultaneously over N time steps. In

order to detect dependencies between X and Y, I utilize two widely used techniques to assess correlation:

1. The Pearson product-moment correlation or correlation coefficient (CC)
2. The Mutual Information (MI).

The correlation coefficient is a normalized measure of the linear correlation between 2 fMRI time-series, whereas the mutual information is an information-theoretic tool used to measure both linear and non-linear dependencies. Explaining the concepts and theory that underlies mutual information would go beyond the scope of this thesis (Cover and Thomas 1991), but it suffices to say that MI measures by how much knowledge of one of the time-series reduces uncertainty about the other time-series. Despite being able to detect virtually any kind of dependency, the power of MI is limited by the fact that the number of samples necessary to obtain a robust estimate increases with the complexity of the dependency between the 2 time-series. The combination of CC and MI methods is an appropriate trade-off between assessing non-linear relationships and obtaining robust estimates of those relationships.

Estimation of the Correlation Coefficient:

For the 2 time-series defined beforehand, the correlation coefficient is given by:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{(E[X^2] - E^2[X])(E[Y^2] - E^2[Y])}}, \quad (1)$$

where $E[.]$ denotes the expected value. Since the expected values are unknown, they need to be estimated to give the following estimate of $\rho_{X,Y}$:

$$\tilde{\rho}_{X,Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \quad (2)$$

where \bar{x} and \bar{y} denote the means of X and Y respectively. It can be shown that $-1 \leq \tilde{\rho}_{X,Y} \leq 1$. Time-series for which $\tilde{\rho}_{X,Y} \approx 0$ are uncorrelated, whereas time-series for which $|\tilde{\rho}_{X,Y}| \approx 1$ have a strong dependency.

Estimation of the Mutual information:

The mutual information between 2 random variables X and Y is expressed in bits of information and is given by:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \quad (3)$$

where $H(X)$ and $H(Y)$ are the entropies of X resp. Y , and $H(X,Y)$ is the joint entropy of X and Y (entropy denoting the amount of uncertainty that is contained in a variable). It can be shown that $0 \leq I(X;Y) \leq \min(H(X), H(Y))$. If $I(X;Y) \approx 0$, the 2 variables are said to be independent; the higher the value of $I(X;Y)$, the stronger the correlation between the 2 variables.

Calculating $I(X;Y)$ requires computing the unknown joint probability distribution of X and Y , but it can be estimated using a frequentist inference (Moddemeijer 1989). For our purposes we consider the 2 time-courses to be discrete realizations of the random variables X and Y . If we partition X into m bins $M_1^X \dots M_m^X$, we can assign a probability p_i^X to each possible outcome M_i^X : $p_i^X = \frac{k_i^X}{N}$ for $i = 1 \dots m$, where k_i^X is the number of times that X falls into bin M_i^X , and N is the length of the time-series. Now the entropy of X can be written as:

$$H(X) = -\sum_{i=1}^m p_i^X \log(p_i^X). \quad (4)$$

If we also partition Y into m bins $M_1^Y \dots M_m^Y$ with respective probabilities $p_j^Y = \frac{k_j^Y}{N}$, we can estimate the joint entropy of X and Y :

$$H(X, Y) = - \sum_{i=1}^m \sum_{j=1}^m p_{ij}^{XY} \log(p_{ij}^{XY}), \quad (5)$$

where $p_{ij}^{XY} = \frac{k_{ij}^{XY}}{N}$ is the joint probability of $X \in M_i^X$ and $Y \in M_j^Y$, and k_{ij}^{XY} is the number of occurrences of the pair (M_i^X, M_j^Y) . Combining (3), (4), and (5) yields the following estimate for the mutual information between X and Y :

$$I_{est}(X; Y) = \sum_{i=1}^m \sum_{j=1}^m p_{ij}^{XY} \log \left(\frac{p_{ij}^{XY}}{p_i^X p_j^Y} \right). \quad (6)$$

However, this estimate is biased due to the finite length of the time-series, the errors introduced by partitioning the time-series into bins, as well as by the fact that the mutual information is bounded by 0. The corrected mutual information I_∞ can be decomposed into an estimate term and a bias term (Roulston 1999):

$$I_\infty(X; Y) = I_{est}(X; Y) + I_{bias}(X; Y), \quad (7)$$

with $I_{bias}(X; Y) = \frac{M_X^* + M_Y^* - B_{MXY}^* - 1}{2N}$, where M_X^* is the number of bins for which $p_i^X \neq 0$, M_Y^* is the number of bins for which $p_j^Y \neq 0$, and M_{XY}^* is the number of bins for which $p_{ij}^{XY} \neq 0$ (Roulston 1999). Note that this bias correction term is always less than or equal to 0. From (6) and (7) the bias corrected estimate of the mutual information between X and Y is thus given by:

$$I_{\infty}(X; Y) = \sum_{i=1}^m \sum_{j=1}^m p_{ij}^{XY} \log \left(\frac{p_{ij}^{XY}}{p_i^X p_j^Y} \right) + \frac{M_X^* + M_Y^* - M_{XY}^* - 1}{2N}. \quad (8)$$

One can notice that in addition to depending on the joint probability distribution of X and Y, this estimate of the mutual information also depends on the binning variable m and the sample size N.

For the purpose of this paper we will consider a constant value for m, but we would like to compare MIs estimated from time-series of different lengths N, and hence we need to normalize the MI estimates. It can be shown that for large values of N (Goebel, Dawy et al. 2005), the distribution of I_{est} follows a Gamma distribution with shape and scale parameters α and β :

$$I_{est} \sim \Gamma(\alpha, \beta) \text{ where } \alpha = \frac{1}{2}(m-1)^2 \text{ and } \beta = \frac{1}{N}, \quad (9)$$

with mean $\bar{I}_{est} = \alpha\beta = \frac{(m-1)^2}{2N}$ and variance $V(I_{est}) = \alpha\beta^2 = \frac{(m-1)^2}{2N^2}$. When adding the bias correction term I_{bias} to I_{est} , some estimates for I_{∞} will be negative. I_{∞} can be shifted back into the positive range by compensating for the largest possible shift (this is achieved when $p_{ij}^{XY} \neq 0, \forall(i, j)$, resulting in $I_{bias} = \frac{m + m - m^2 - 1}{2N} = \frac{-(m-1)^2}{2N}$). It can be shown empirically through simulation that a shifted I_{∞} still follows a Gamma distribution:

$$I_{\infty} \sim \Gamma\left(\alpha', \frac{c}{N}\right), \quad (10)$$

where c is some constant, and α' is independent of N.

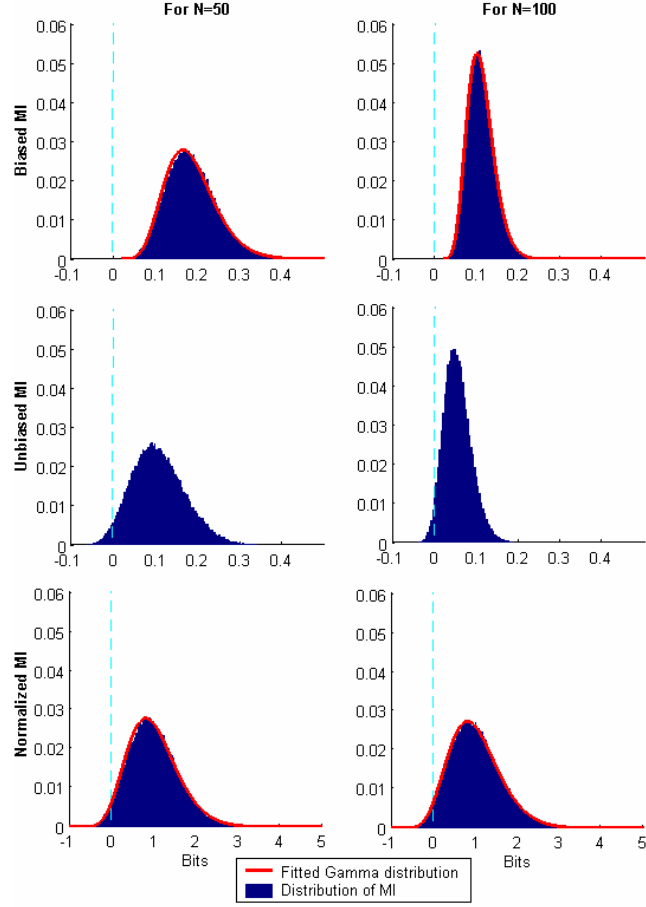


Fig. 31: **Distribution of the Mutual Information.** The mutual information (MI) between 2 random time-courses X and Y of length N (composed of scrambled fMRI signal magnitudes) was calculated using different estimator functions for MI. This procedure was repeated 100,000 times to obtain the distribution of the mutual information. Panel A shows the histogram of the biased distribution I_{est} for $N=50$ (left column) and $N=100$ (right column), as well as the best-fitting Gamma distribution. When applying the bias correction, the new distribution I_{∞} shifts to the left (Panel B), but the shape remains the same. After normalizing the distributions (I_{∞}^{norm}), the MIs for $N=50$ and $N=100$ follow the same Gamma distribution (Panel C), and can thus be directly compared.

Now, by using the following property of the Gamma distribution:

$$\text{if } X \sim \Gamma(a, b) \text{ then } t\Gamma(a, b) \sim \Gamma(a, tb), \forall t > 0, \quad (11)$$

and the mean value of I_{∞} , it can be shown that:

$$I_{\infty}^{norm} = \frac{I_{\infty}}{\bar{I}_{\infty}} \sim \Gamma\left(\alpha', \frac{1}{\alpha'}\right), \quad (12)$$

which is independent of N . Thus all I_{∞}^{norm} are comparable for any value of N , i.e., MI estimates for time-series of different lengths can be directly compared after normalizing the different estimates according to (12). The normalization procedure is illustrated in Figure 31 for two values of N : $N = 50$ (left column) and $N = 100$ (right column). Panel A shows the distribution of I_{est} , panel B shows the shifted distribution of I_{∞} , and panel C shows the normalized distribution I_{∞}^{norm} .

III.5.3. Methods

fMRI Statistical Analysis

A general linear model (GLM) analysis was performed in SPM2 to identify brain regions whose BOLD response reflected information revealed by the Investment and Repayment screens. First a separate general linear model for each subject was specified (fixed effects analysis). All images were high-pass filtered in the temporal domain (filter width 128 s) and autocorrelation of the hemodynamic responses was modeled as an AR(1) process. In the GLM model all visual stimuli and motor responses were entered as separate regressors that were constructed by convolving a hemodynamic response function (hrf) with a comb of Dirac functions at the onset of each visual stimulus or motor response. The two regressors of interest were β_{inv} and β_{rep} , corresponding to the revelation of the investment and repayment screens respectively. Specifically, a one-sample t-test on β_{inv} and β_{rep} (random effects analysis) was performed to detect voxels where $\beta_{inv} > 0$ and $\beta_{rep} > 0$, i.e., a test that looked for regions of interest (ROIs) that were significantly more activated during the revelation of the investment and repayment screens in both the Investor and Trustee brains.

fMRI Time-Series Analysis

We extracted and averaged the time-courses of the 27 voxels surrounding the peak of activation of each ROI, and removed the low-frequency components with a running-average filter of length 21. In order to compare time-courses acquired from different subjects and on different MRI scanners, we demeaned and normalized the time-courses, and synchronized them to the events of interest. We verified that the distribution of the magnitudes followed a standard normal distribution with mean 0 and variance $\sigma = 1$.

We next calculated the correlation between selected time points from the normalized time-courses using the two aforementioned dependency measures—the correlation coefficient CC and the mutual information MI. The CC was calculated using the formula in (2). For MI each time-series was quantized into $m=6$ bins, where the binning limits were $[-\infty; -2\sigma; -\sigma; 0; +\sigma; +2\sigma; +\infty]$ with $\sigma = 1$. The mutual information was calculated according to (8) and normalized according to (12), in order to obtain a percentage change of MI.

III.5.4. Results

fMRI Results

The structure of the finitely repeated trust game requires subjects to modify their behavior as the game evolves and as information about their opponents' strategies becomes available. Although players decide how much money to invest or repay during the decision phase, we did not focus on that part of the game, as it is difficult to determine when exactly players make their decision. There are however two critical moments in each round that are at the core of strategic choice: the revelation of the results from the investment and repayment phases. As this new information is a basis for subsequent behavior and the resulting game dynamics, we performed a whole-brain analysis to detect areas that showed a high BOLD response during both the investment and repayment phases. We used a

general linear model, where the 2 regressors of interest were β_{inv} and β_{rep} , corresponding to the revelation of the investment and repayment screens respectively.

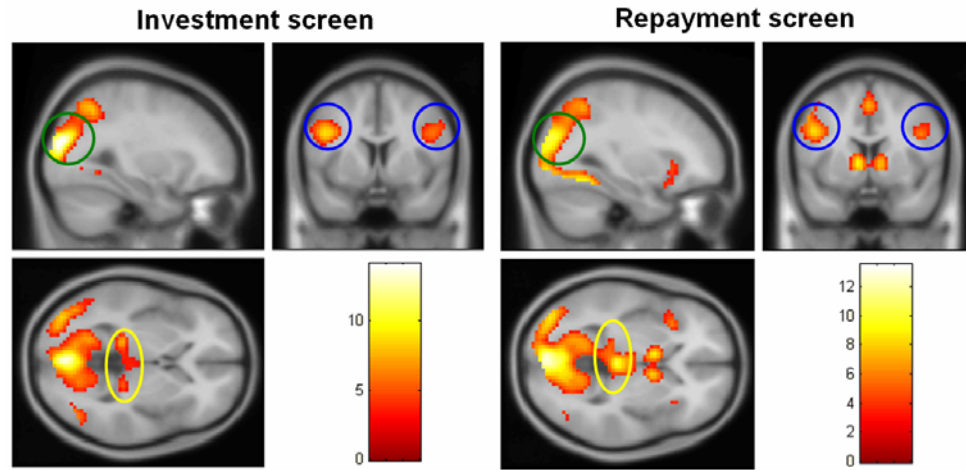


Fig. 32: **Activations in the Investor Brain.** Activations in the Investor brain at the moment where the results of the investment phase (left panel) and repayment phase (right panel) are revealed ($p \leq 0.001$, minimum cluster size: 5, $x = -30$, $y = 6$, $z = 0$). Areas that are activated under both conditions are highlighted: occipital gyrus (green circles), inferior frontal gyrus (blue circles), and midbrain (yellow circles).

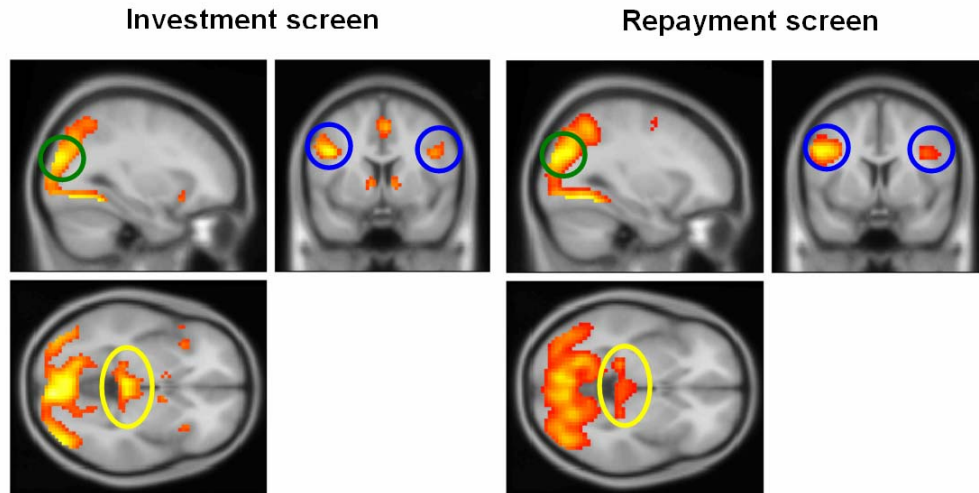


Fig. 33: **Activations in the Trustee Brain.** Activations in the Trustee brain at the moment where the results of the investment phase (left panel) and repayment phase (right panel) are revealed ($p \leq 0.001$, minimum cluster size: 5, $x = -30$, $y = 6$, $z = -3$). Areas are highlighted in the same way as in the previous figure.

Three regions that were active during both revelation screens in Investors and Trustees were identified: middle/superior occipital gyrus (BA19), inferior frontal gyrus (bilateral BA9), and medial and lateral midbrain (Fig. 32 & 33, Table 3). These three regions were used as ROIs for the subsequent analysis. Note that some other areas were also activated during one of the revelation screens: bilateral caudate; superior frontal gyrus (BA6/8), bilateral orbitofrontal cortex (BA47) and medial frontal gyrus (BA11/32), but as none were activated for both screens, they were excluded from the analysis. Several other areas in the occipital lobe were also ignored as they were just activations in response to the visual stimuli.

Region (BA)	Coordinates			Investor Inv. screen	Investor Rep. screen	Trustee Inv. screen	Trustee Rep. screen
	X	y	z				
Occipital Gyrus (BA19)	-30	-81	24	14.06	10.02	7.37	9.64
	30	-81	24	13.24	10.66	7.68	11.87
Inf. Fr. Gyr. (BA9)	-45	6	30	8.60	8.24	7.82	9.71
	45	6	30	5.61	5.38	4.88	4.11
Midbrain	-21	-30	-3	8.10	6.60	4.14	6.40
	21	-30	-3	6.37	5.76	3.76	5.08
	0	-27	-3	4.72	13.47	7.16	4.86

Table 3: **Summary of activations at the revelation screens.** This table shows the t-values of areas that are activated in both Investors (46 subjects) and Trustees (47 subjects) at the moment of the revelation of the Investment and Repayment screens.

Game Dynamics

To detect temporal (i.e., dynamic) dependencies within areas, we needed to quantify strategic behavior in the trust game. As both Investor and Trustee made decisions in every round, they continuously had to adapt their strategy in response to what the other player did, resulting in a rich behavioral interaction. Consequently, we examined how changes in repayment ratios ΔR_i were being made as a result to changes in investment ratios ΔI_i . The change in investment ratio in round i is defined as $\Delta I_i = I_i / 20 - I_{i-1} / 20$, where I_i and R_i are the investment and repayment amounts respectively for that round. Similarly, the

change in repayment ratio is defined as $\Delta R_i = R_i / (3I_i) - R_{i-1} / (3I_{i-1})$. This representation segregated the behavioral space into four quadrants, each reflecting strategies that the Investor and Trustee could follow (see Fig. 26). From a game-theoretic perspective, reciprocal events (green quadrants) reflect tit-for-tat strategies, which are a robust way to create human cooperation in repeated games (Axelrod and Hamilton 1981). Non-reciprocal events (red quadrants) are the result of altruistic (benevolent) and greedy (malevolent) strategies. Neutral events (blue circle) occur when both players have reached some stable pattern of exchange.

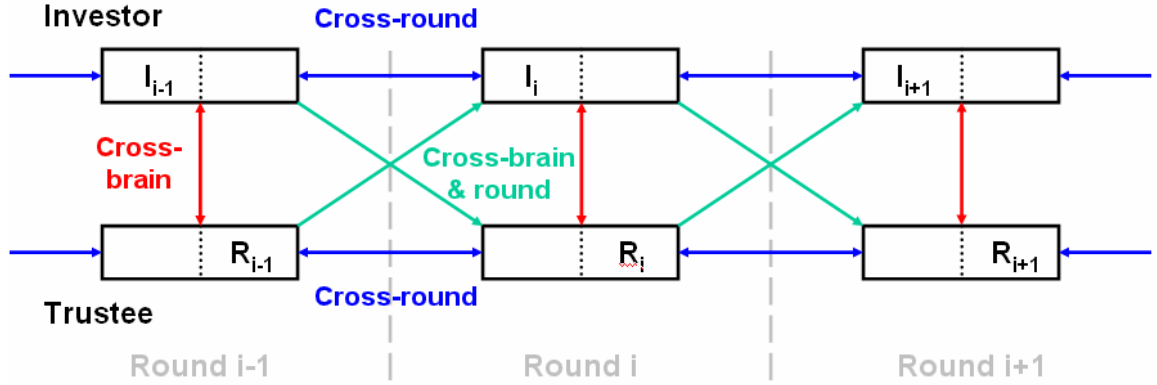


Fig. 34: **Game Dynamics.** Schematic of the different possible interactions between brain areas in the trust game: cross-round interactions, cross-brain interactions, and cross-brain-and-round interactions. I_i and R_i represent the invested and returned money amounts, respectively, in round i .

Next, we related these three types of behavioral strategies (reciprocal, non-reciprocal, and neutral) to the following game dynamics (Fig. 2B):

1. Cross-round interactions: dependencies between time-courses at round $i-1$ and round i for the Investor (CR_i^I) and the Trustee (CR_i^T),
2. Cross-brain interactions: dependencies between Investor and Trustee time-courses in round i (CB_i),

3. Cross-brain-and-round interactions: dependencies between Investor time-courses in round $i-1$ and Trustee time-courses in round i (CBR_i^{TT}) and dependencies between Trustee time-courses in round $i-1$ and Investor time-courses in round i (CBR_i^{TI}).

Note that for the cross-brain interactions we calculated the mutual information (MI) and the correlation coefficient (CC) between the time-series from the entire round, whereas for the other types of interactions we calculated MI and CC from the concatenated time-courses of the investment and repayment screens.

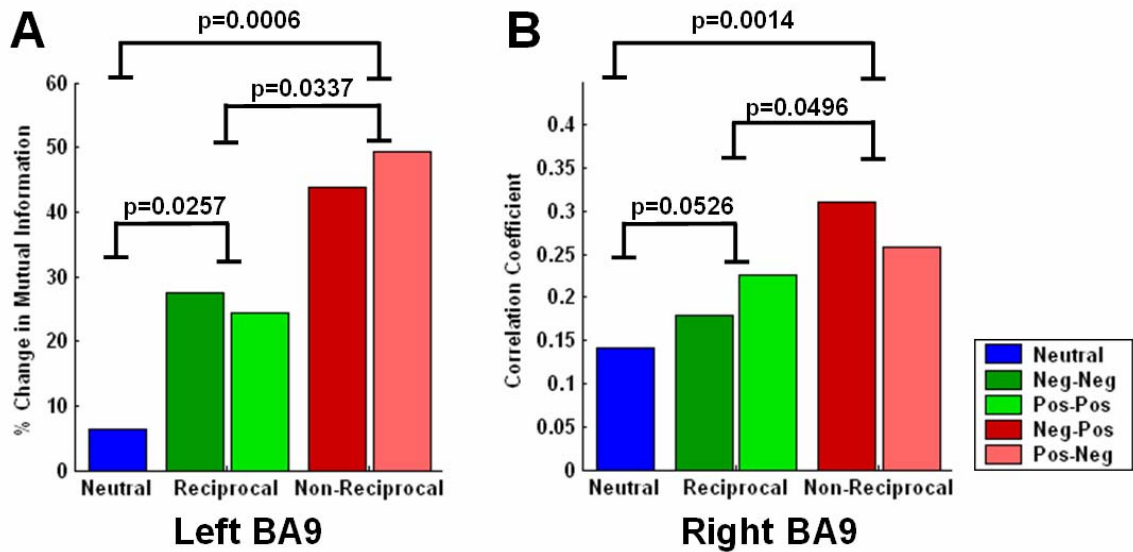


Figure 35: **Cross-Round Dynamics in BA9.** Panel A shows how the mutual information between time-courses from successive rounds segregates among between 3 different strategic types of behavior (neutral, reciprocal, and non-reciprocal) in the left BA9. Reciprocal events include ‘pos-pos’ events (increase in the investment ratio followed by an increase in the repayment ratio) and ‘neg-neg’ events (decrease in investment ratio followed by a decrease in the repayment ratio). Non-reciprocal events include ‘pos-neg’ (increase in the investment ratio followed by a decrease in the repayment ratio) and ‘neg-pos’ (decrease in investment ratio followed by an increase in the repayment ratio). Neutral events occur when neither the investment nor the repayment ratio changes. The mutual information (MI) associated with non-reciprocal events is significantly larger than the MI associated with reciprocal events, which in turn is significantly larger than the MI associated with neutral events. Panel B shows how the correlation coefficient segregates strategies in the same way in the right BA9. All p values are obtained from 1-sided t tests.

We found that the dependencies could be segregated into the 3 behavioral categories defined above, e.g., Investor cross-round interactions in the left BA9 (as calculated by mutual information) are the largest for non-reciprocal events, and the lowest for neutral events (Fig. 35A). A similar pattern is found in the right BA9 when using the correlation coefficient instead of the mutual information (Fig. 35B). This trend can be observed for all 3 types of interactions in different ROI's, and is summarized in Table 4. In some cases only the MI would discriminate between categories, and for other cases only the CC would do so, showing that a combination of both methods is necessary to detect differences.

Game Dynamics	Area	DM	N	R	NR	N<R	N<NR	R<NR
Investor Cross-round	BA9 L	MI	.063	.256	.471	0.0337	0.0006	0.0257
	BA9 R	CC	.141	.208	.280	0.0526	0.0014	0.0496
	BA19 L	CC	.319	.393	.488	0.0271	<0.0001	0.0059
	BA19 R	CC	.259	.361	.402	0.0038	0.0009	/
Trustee Cross-round	BA9 L	CC	.210	.255	.333	/	0.0032	0.0399
	BA9 R	CC	.137	.202	.245	0.0499	0.0032	/
Cross-brain	Midbr. R	MI	.065	.242	.365	0.0168	0.0013	0.1113
Cross-brain-and-round CBR Investor-Trustee	BA19 L	CC	.079	.182	.207	0.0022	0.0010	/
	BA19 R	CC	.042	.195	.246	<0.0001	<0.0001	/
Cross-brain-and-round CBR Trustee-Investor	BA9 L	CC	-0.03	.036	.136	0.0280	<0.0001	0.0053
	BA9 R	MI	-0.14	0	.241	0.0526	0.0002	0.0068
	BA19 L	CC	.088	.124	.253	/	<0.0001	0.0018
	BA19 R	CC	.091	.142	.246	.0828	0.0002	0.0106

Table 4: **Summary of game dynamics.** DM denotes what dependency measure was used: mutual information (MI) or correlation coefficient (CC). The magnitudes of the MI or CC associated with neutral (N), reciprocal (R), and non-reciprocal (NR) are listed in columns 4–6. Columns 7–9 show the statistical significance of the relationship $N < R < NR$ using 1-sided t-tests.

III.5.5. Discussion and Conclusion

In this study we have shown how mutual information and correlation coefficients can be used to detect temporal dependencies in fMRI brain responses and relate them to behavioral interactions. Those interactions, labeled as game dynamics, occur across rounds and across the brains of interacting players.

More specifically, cross-round interactions in Investors (CR^I) and Trustees (CR^T), as well as cross-brain-and-round interactions between Trustees and Investors (CBR^{IT}) segregated between neutral, reciprocal, and non-reciprocal strategies in bilateral BA9. A similar division between strategies was observed in bilateral BA19 for cross-round interactions in Investors (CR^I) as well as for cross-brain-and-round interactions (both CBR^{IT} and CBR^{TI}). Cross-brain (CB) interactions between Investors and Trustees also segregated between the 3 types of strategies in the midbrain. In all types of interactions the nature of the segregation was the same, with non-reciprocal events having the largest level of dependency, and neutral events having the lowest level of dependency. In Section III.4. we reported a similar split-up between time-courses in the Trustee caudate and midbrain which actually predicted the strategy chosen by the Trustee. Although the magnitude ranking for the three types of strategies is slightly different in this study, this confirms the hypothesis that the brain treats them differently.

We next investigated why the magnitude of the MI/CC is different for the three types of strategy and how it relates to the behavior. Since activation was found in BA9, the first hypothesis was that it is directly related to the visual display of the revelation screens. Intuitively, if 2 screens are visually very similar, then the time-courses associated with those screens should also be very similar, resulting in a high MI/CC. Conversely, if the 2 screens are visually very different, then the time-courses should also be very different, resulting in a low MI/CC. Now, when players have converged to some kind of equilibrium in the trust game, i.e., when there is no change in investment or repayment ratio, then the MI/CC between successive screens should be very high, as the screens will look exactly the same. This is however exactly to opposite of what our data suggests with neutral strategies

displaying the lowest MI/CC. We performed an additional control by extracting the time-courses from an area in the primary visual cortex (BA17), which was also activated during the reveal screens. When segregating the MI/CC into the 3 categories, no difference could be observed (Fig. 36). As this suggests that differences in MI/CC cannot be related to visual stimuli, we investigated other possible hypotheses and focused first on the cross-round (CR) and cross-brain-and-round interactions (CBR).

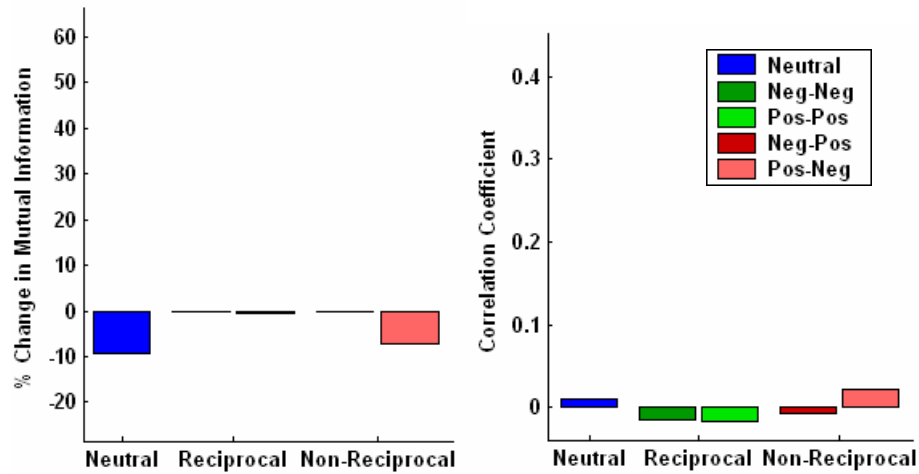


Fig. 36: **Cross-round dynamics in primary visual cortex.** (BA17, $x=-12$, $y=90$, $z=0$). This control analysis shows that neither the mutual information (left panel) nor the correlation coefficient (right panel) between time-courses from successive rounds segregates among strategic behaviors. All differences between neutral, reciprocal, and non-reciprocal categories are non significant (t-test).

CR and CBR interactions occur in two areas, namely the middle/superior occipital gyrus (BA19) and the inferior frontal gyrus (BA9). BA9 is part of the dorsolateral prefrontal cortex (DLPFC), and the actual activation in BA19 is at the border between the cuneus and the precuneus. All three structures are involved in memory processing. More specifically, both the cuneus (Andreasen, O'Leary et al. 1995; Cabeza, Grady et al. 1997; Otten and Rugg 2001) and the precuneus (for a review, see Cavanna (Cavanna and Trimble 2006)) have been shown to play a role in memory retrieval, and the DLPFC is involved in working memory processes (Petrides 1994; Goldman-Rakic 1996) and in memory encoding and retrieval (Sandrini, Cappa et al. 2003). Moreover, the precuneus and prefrontal cortex have

been shown to be strongly interconnected and these projections tend to concentrate at the level of BA 8, 9, and 46 (Cavanna and Trimble 2006). We thus hypothesize that the differential level of MI/CC during CR and CBR interactions is due to the level of intensity of memory processing and retrieval in those areas. More specifically, the neutral strategy does not require much memory processing since the investment and repayment ratios have not changed from one round to another. For the reciprocal strategy, the brain needs to retrieve the information from the previous round and process it in order to evaluate the change in increase in investment and repayment ratios, resulting in a higher dependency between rounds. Finally, if the subjects are playing a non-reciprocal strategy, the memory load is even higher as the brain needs to assess not only the magnitude of the change, but also the direction of the change. This claim is supported by two recent studies: in an economic decision-making study (Deppe, Schwindt et al. 2005) it has been shown that making an easy decision (picking a favorite brand vs. picking another brand) reduces activity in the cuneus/precuneus as well as BA9. In another study (van Leijenhorst, Crone et al. 2006) both of those areas showed increased activity for high vs. low risk trials. Although both of those studies were decision-making studies, and were not specifically designed to look at memory load, one can argue that a hard or risky decision requires more memory load as opposed to an easy one, which is in line with our findings.

The segregation between behavioral strategies was also found in the midbrain for cross-brain (CB) interactions. Although the midbrain is not implicated in any memory retrieval tasks, a similar explanation about the correlation magnitudes can be advanced. Whereas the CC/MI interactions were obtained for the revelation screens, the cross-brain interactions were calculated over the course of the whole round, resulting in stronger MI for non-reciprocal events and weaker MI for neutral events. The midbrain dopaminergic system plays an important role in the brain reward system (Schultz, Dayan et al. 1997) and has been shown to make predictions about likely rewards. We hypothesize that the reward computation between brains of two interacting players is more intensive and thus more synchronized for the non-reciprocal strategy than for the reciprocal or neutral strategies, resulting in higher mutual information.

We would like to emphasize the fact that in addition to finding temporal interactions within a brain area, we also detected interactions between the brains of interacting players in BA9, BA19, and midbrain. In Section III.6. it is shown how a common trust signal shifts backward in time in the caudate as the two players learn about each other, but the present findings are the first published results that show how the brains of two interacting players work together when exposed to various strategies. With respect to this, the usage of the mutual information as an alternate tool to the correlation coefficient turned out to be extremely useful, as it was able to pick up additional interactions. In Section III.5.2. we developed and standardized a method to use mutual information as a tool to capture non-linear brain responses. Although mutual information is able to detect virtually every kind of correlation, the drawback is that the amount of data points needed to detect a dependency increases with the complexity of the correlation. This justifies the combined use of the mutual information and the correlation coefficient to detect dependencies in the brain, which has been shown to be a source of non-linearity. We would also like to call attention to the fact that we are not making any claim about the nature of the interaction of the brain, but merely detecting that there is a differential interaction based on the used strategy. As the actual interaction might vary from subject to subject, it would be incorrect to average over all subjects, and the limited number of events per trust game does not allow me to identify individual differences. This is however something that can be quite easily investigated if one disposes of enough trials by looking at the joint probability distribution of the signal magnitudes of two fMRI time-series.

Lastly, it should also be noted that we did not look for possible interactions between different areas, but limited ourselves to temporal dependencies between identical brain structures and between players. When investigating the interactions between structures, special attention has to be devoted to the time component, as different structures might be activated at different times. This problem seems to be more easily approached by using methods from functional connectivity studies (Friston 2003; David, Cosmelli et al. 2004).

III.6. Trust and Reciprocity

Note: The work in this section was mostly done by Brooks King-Casas, and is included here for completeness as it uses the same data. It was published as "Getting to know you: Reputation and trust in a two-person economic exchange." by King-Casas, B., Tomlin, D., Anen, C., Camerer, C., Quartz, S. and Montague, P.R. in *Science* **308**(5718):78-83 (2005).

III.6.1. Background

The expression and repayment of trust is an important social signaling mechanism that influences competitive and cooperative behavior (Trivers 1971; Axelrod and Hamilton 1981; Coleman 1990; Rachlin 2002; Adolphs 2003; Fehr and Fischbacher 2003). The idea of trust typically conjures images of complex human relationships, so it would seem to be a difficult part of social cognition to probe rigorously in a scientific experiment. Nevertheless, instances of trust can be stripped of complicating contextual features and encoded into economic exchange games that preserve its essential features (Camerer and Weigelt 1988; Fehr, Kirchsteiger et al. 1993; Berg, Dickhaut et al. 1995). For example, in a game in which two players send money back and forth with risk, trust is operationalized as the amount of money a sender gives to a receiver without external enforcement (Berg, Dickhaut et al. 1995). Such trust games now enjoy widespread use both in experimental economics (Camerer 2003) and neuroscience experiments (McCabe, Houser et al. 2001; Rilling, Gutman et al. 2002; Eisenberger, Lieberman et al. 2003; Sanfey, Rilling et al. 2003; de Quervain, Fischbacher et al. 2004; Decety, Jackson et al. 2004; Glimcher and Rustichini 2004). By using a multi-round version of the trust game, (i) trust becomes bidirectional, in that both the investor and trustee assume the risk that money sent might not be reciprocated by their partner; and (ii) reputation building can be probed, as players develop models of one another through iterated exchange.

III.6.2. Reciprocity Predicts Trust

Linear regression analyses of the behavior of 48 pairs of subjects identified reciprocity to be the strongest predictor of subsequent increases or decreases in trust (see Section III.6.6. for methods). Reciprocity is defined as a fractional change in money sent across rounds by one player in response to a fractional change in money sent by their partner. This definition is simply an operationalized version of tit-for-tat, that is, a repayment in kind. Deviations from neutral reciprocity (perfect tit-for-tat) act as a strong social signal in the context of this game.

In particular, strong deviation in investor reciprocity was the best predictor of changes in partner trust and became the primary focus of our analysis. Investor reciprocity on round j was quantified as $\Delta I_j - \Delta R_{j-1}$, where ΔI_j is the fractional change in investment from round $j-1$ to j and ΔR_{j-1} is the last fractional change repayment ($R_{j-1} - R_{j-2}$).

We divided the exchanges of the 48 subject pairs into three approximately equal-sized groups: (i) benevolent reciprocity, (ii) neutral reciprocity, and (iii) malevolent reciprocity. These behavioral exchange data are summarized in Fig. 37A. For benevolent reciprocity, investors are actually being generous (sending more) in response to a defection by the trustee (decrease in repayment) (left panel). Conversely, for malevolent reciprocity, the investor repays the trustee's generosity with a breach of trust (right panel).

Using a general linear model analysis, we first sought trustee brain regions whose blood oxygenation level-dependent (BOLD) response was greater for benevolent or malevolent investor reciprocity than for neutral investor reciprocity. This analysis identified four significant regions: inferior frontal sulcus, superior frontal sulcus, thalamus, and inferior/superior colliculli. These findings are consistent with a surprise signal—an unsigned response to deviations in the expected behavior of one's partner. A second analysis, comparing BOLD response for benevolent reciprocity to BOLD response for malevolent reciprocity, identified significant differences only in the head of the caudate nucleus (Fig. 37B and C): (i) BOLD response was greater for instances of benevolent

reciprocity relative to malevolent and neutral reciprocity; and (ii) responses to malevolent reciprocity did not differ from those to neutral reciprocity. These voxels were subsequently subjected to a region-of-interest (ROI) analysis.

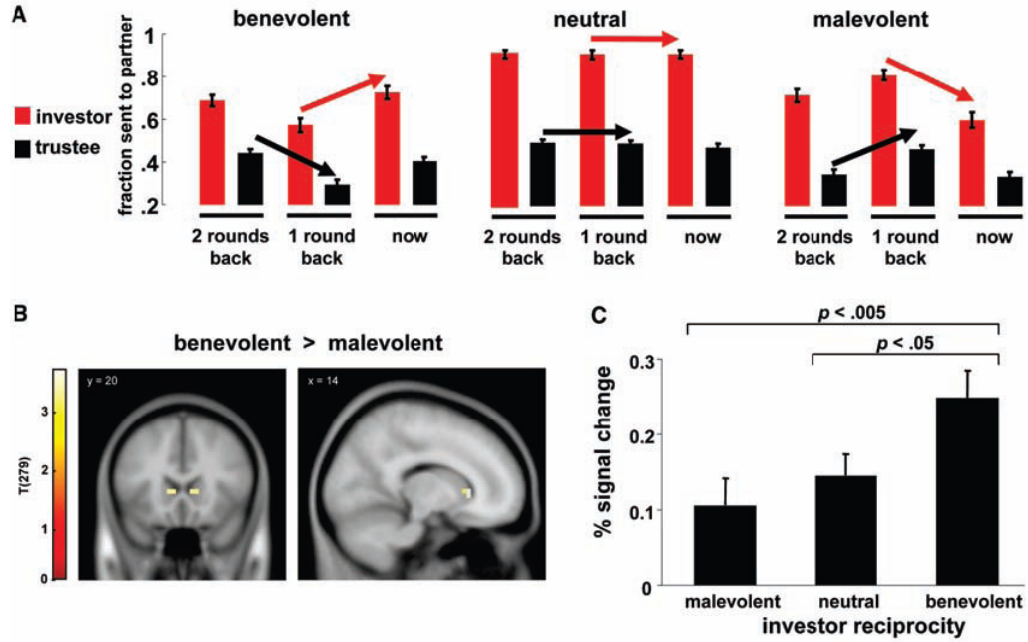


Fig. 37: Correlates of reciprocity in a multi-round economic exchange. (A) Behavioral summary. Mean \pm SE of investor (ΔI , red) and trustee (ΔR , black) behavior of rounds contributing to benevolent ($n=125$), neutral ($n=134$), and malevolent ($n=125$) investor reciprocity categories. In each round j , investor reciprocity was defined as $r_j = \Delta I_j - \Delta R_{j-1}$; that is, the difference between the current change in payment ΔI_j by the investor in response to the previous change in repayment ΔR_{j-1} by the trustee. In the case of benevolent reciprocity, investors are being generous (sending more) in response to a defection by the trustee (decrease in repayment). Likewise, in the case of malevolent reciprocity, the investor repays the trustee's generosity (increase in previous repayment) with a breach of trust. (B) Response of trustee brain to investor reciprocity. A general linear model analysis identified four regions in the trustee brain that showed responses that were greater for the revelation of malevolent and benevolent investor reciprocity than for neutral reciprocity. Only one region, the head of the caudate nucleus, showed a response that was greater for benevolent relative to malevolent reciprocity (statistical parametric map shown alongside pseudo-color legend). No region showed greater responses to malevolent relative to benevolent investor reciprocity. (C) Region-of-interest analysis of head of caudate in trustee brain. Average activity 6 to 10 s after the investor's decision is revealed to trustee shows that the brain response to benevolent reciprocity was significantly greater from neutral (two-tailed t test, $p < 0.05$) and malevolent reciprocity (two-tailed t test, $p < 0.005$).

III.6.3. “Intention to Trust” Signals

We expected to find a hemodynamic response in this ROI that correlated with the trustee’s next choice to repay, and we expected that such signals might show strong cross-brain correlations. The reason for this expectation derived from the fact that reciprocity expressed by the investor ($\Delta I_j - \Delta R_{j-1}$) strongly predicted ($r=0.56$) future changes in trust repayment, ΔR_j) by the trustee. For example, benevolent reciprocity by the investor is expected to generate the intention to increase repayment (trust) in the brain of the trustee. A similar intention to decrease trust (repayment) would be expected in the trustee brain following malevolent reciprocity by the investor. Some part of the investor’s brain should anticipate the neural consequences of changes in their own reciprocity on the trustee’s brain; therefore, we also expected that such “intention to trust” signals would show strong cross-brain correlations. Indeed, they did.

III.6.4. Model Building of Partner: Cross-Brain Analysis

To carry out this analysis, we separated the hemodynamic responses in the caudate of the trustee brain into three groups according to whether their next repayment was larger, smaller, or the same as their last repayment. We were particularly interested in the net neural response to the intention to increase trust (repayment), because this act embodies risk on the part of the trustee and signals to the investor a degree of willingness to cooperate. We computed the net intent-to-trust signal in the ROI of the trustee caudate as:

$$H(\text{increased repayment next round}) - H(\text{decreased repayment next round})$$

where H represents the hemodynamic response. Using this difference signal in the trustee brain, we computed cross-brain correlations with the investor brain and sought regions with the largest correlations. We were particularly interested in how the cross-brain correlations might change as the task developed and the subjects built better models of one another.

Consequently, changes in this signal were examined across early (3 and 4), middle (5 and 6), and late (7 and 8) rounds using crossbrain and within-brain correlational analysis. Figure 38 illustrates the cross-correlograms of this signal with activity in two regions: the

middle cingulate cortex (MCC) of investors and the anterior cingulate cortex (ACC) of trustees. The blue traces indicate that MCC activity in the investor brain and ACC activity in the trustee brain were most strongly correlated ($r > 0.59$) when the MCC signal was shifted forward in time by 14 s. The important point here is that the strongest cross-brain correlation did not shift significantly in time from early to late rounds; that is, neural responses in both brains to fiducial markers of the task did not change relative to each other. However, the peak of the cross-correlogram between investor MCC activity and the trustee “intention to trust” signal in the caudate showed a pronounced 14 s shift from early to late rounds (green traces). A similar finding resulted for the within-brain analysis of the trustee, using ACC activity and the same “intention to trust” signal in the caudate (red traces). These analyses show that a dramatic change in the relative timing of the measured BOLD signals was taking place either in the “intention to trust” signal of the trustee caudate or in both the trustee ACC and investor MCC. As shown in Fig. 39, the source of the shift is in the “intention to trust” signal of the trustee caudate.

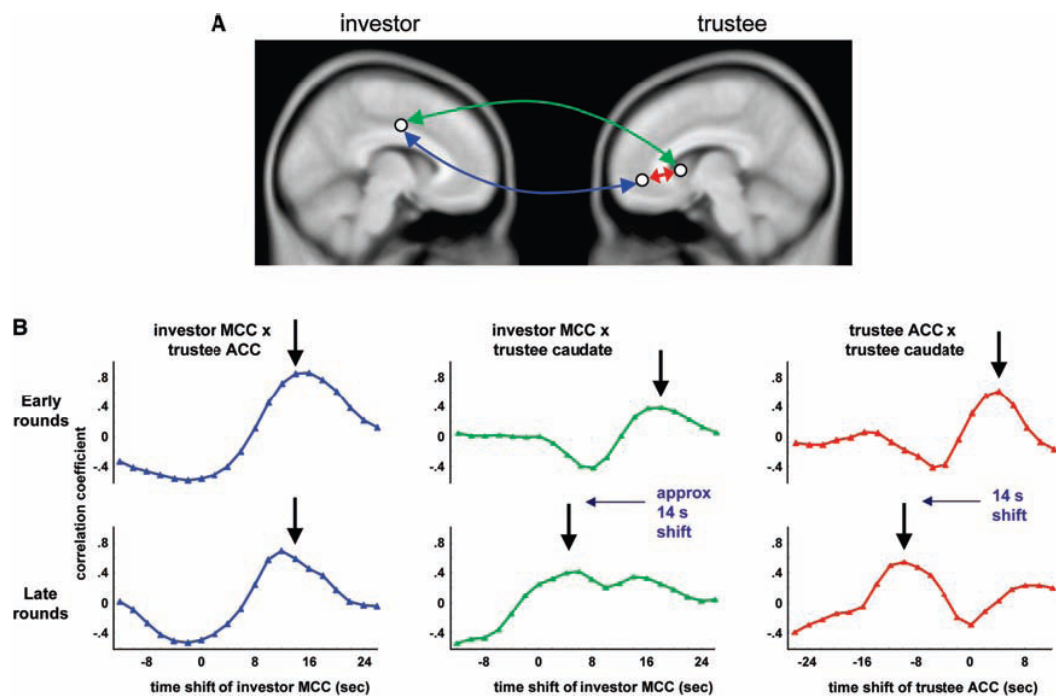


Figure 38: Correlograms of the “intention to trust” with activity in investor MCC and trustee ACC. (A) Regions of correlation. The “intention to trust” signal in the trustee caudate was correlated within and between-brains with regions that responded strongly to

basic behavioral events within each round: The middle cingulate cortex (MCC) of the investor was strongly active when the investor lodged a decision, and the anterior cingulate cortex (ACC) of the trustee was strongly activated when an investor's decision was revealed. (B) Correlograms of caudate, ACC, and MCC. The caudate signal between rounds of increased and decreased repayment isolated an "intention to trust" signal in trustees. Average "intention to trust" signal was correlated with average ACC signal of trustee and average MCC signal of investors during the investment phase of each round and is plotted with different time shifts. Correlograms are shown for early (rounds 3 and 4) and late (rounds 7 and 8) periods of the game. Blue traces indicate that the strongest cross-brain correlation for responses to basic behavioral events of the game did not shift significantly in time from early rounds to late rounds. The peak of the crosscorrelogram between investor MCC activity and the trustee "intention to trust" signal in the caudate shows a pronounced 14 s shift from early to late rounds (green traces). A similar result is evident in the within-brain analysis of the trustee, using ACC activity and the same signal in the caudate (red traces).

Figure 39 shows the time traces of the hemodynamic responses in the head of the trustee caudate segregated according to future changes in trust (increases are shown in black, decreases in red). The amplitude and time effects associated with the 14 s time shift are shown in Fig. 39A and summarized in the bar graphs in Fig. 39B. In early rounds of the task (rounds 3 and 4), the peak of the response for intended increases in trust (i.e., an increase in next repayment) occurs after the investor's decision is revealed. In middle rounds (rounds 5 and 6), this response begins to drop back toward baseline and begins to grow at a time just before the revelation of the investor's decision. By late rounds (rounds 7 and 8), this peak is anticipatory and occurs before the revelation of the investor's decision. These data are consistent with a signal for intended increases in trust changing from being reactive to anticipatory and suggest that the trustee is building a model of the investor's likely next move. To test this model-building idea directly, we performed a separate version of the trust game and queried the trustees on each round about their expectation of the next investment.

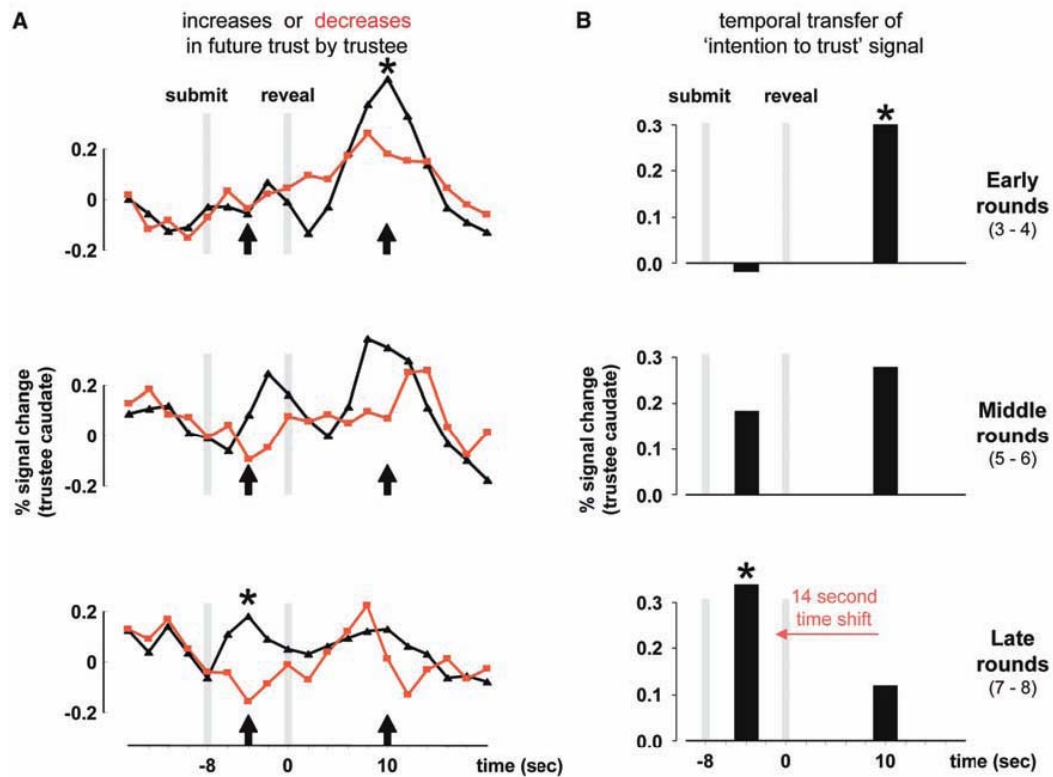


Fig. 39: Neural correlates of reputation building in the trustee brain. (A) ROI time series. An ROI analysis was performed on voxels identified by the contrast illustrated in Fig. 37B. We segregated hemodynamic responses in response to the revelation of the investment (time=0s) according to the next decision made by the trustee (trustee's decision period begins at t= 22s). Hemodynamic amplitudes for future increases in trust ($\Delta R > 5\%$; black trace) were greater ($p < 0.05$) than future decreases in trust ($\Delta R < -5\%$; red trace) in early rounds (top). As the game progressed (middle and bottom), the peak of this differentiated response underwent a temporal transfer from a time after the revelation of the investor's decision (t=10 s; a reactive signal) to a time before this same revelation (t=-4 s; an anticipatory signal). Traces represent subsamples of 144 rounds in which repayment increased or decreased $\geq 5\%$ (mean=20; SD=4.4). (B) ROI bar plot. The difference between the intention to increase trust (black trace of (A)) and the intention to decrease trust (red trace of (A)) is plotted for t=-4 s and t=10 s. The 14 s temporal transfer from reactive to anticipatory is consistent with the development of a reputation for the investor within the trustee brain.

Figure 40 illustrates the results of this additional experiment (n=21 pairs, behavior only). On each round, both the investor and trustee were simultaneously prompted. The investor was cued to make their investment and the trustee was cued to guess the investor's decision (Fig. 40A). Timings were otherwise kept the same. The results of these experiments are summarized as the fraction of highly accurate guesses (to within $\pm \$1$) by the trustee as a

function of round. Notice that the increase in the trustee's accuracy across rounds parallels the time during which the temporal transfer of the neural signal correlated with future increases in trust.

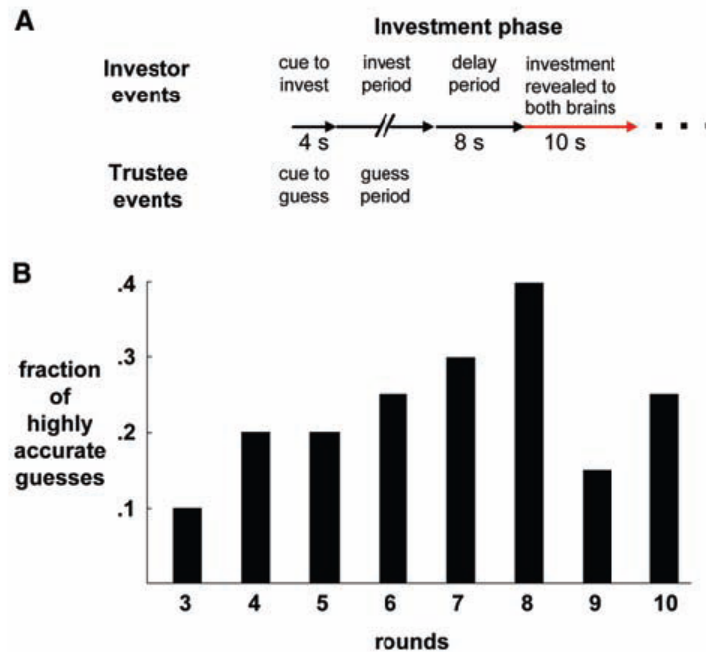


Fig. 40: **Model building in the trustee brain.** In a separate anonymous trust game ($n=21$ pairs), trustees were queried to “guess the amount invested” just before the revelation of the investor’s payment decision to both brains; otherwise, the task was identical to that of the original game ($n=48$ pairs) from which scanning data were derived. (A) Timeline for queries to each player (investor and trustee). During the investment phase of the exchange, the trustees were prompted to guess the investor’s decision. The trustee response to this query was not revealed to the investor. (B) Model building—highly accurate guesses by trustee of investor’s next payment. A highly accurate guess was defined as ± 1 monetary unit from the actual investment ($\pm 5\%$). These data show that a model of the investor’s next move is available to the trustee by the middle to late rounds of the exchange and is not available in the early rounds.

III.6.5. Discussion and Conclusion

We used an anonymous trust game in conjunction with event-related fMRI to probe neural correlates of the expression and repayment of trust between interacting human subjects. Important social relationships are rarely a single expression of trust between two strangers; thus, we made the game multi-round instead of one-shot. Specifically, we sought to

examine trust in a context in which (i) trust was expressed by both partners in the relationship, and (ii) trust could change over time and with experience (Dayan and Abbott 2001).

Using a multi-round trust game and a large sample of subjects ($n=48$ pairs), we identified a social signal (reciprocity) expressed by the investor that strongly predicted changes in trust by the trustee. This social signal elicited two notable effects in the trustee brain: (i) brain regions whose activity correlated with large changes in reciprocity in a manner consistent with a surprise response; and (ii) a specific brain region, the head of the caudate nucleus, where the BOLD response was greater for benevolent reciprocity than for malevolent reciprocity. The strong relation between investor reciprocity and subsequent changes in trustee repayment led us to probe the “intention to trust” in the caudate nucleus. Rounds were segregated on the basis of whether trustees subsequently increased or decreased their repayment, representing a signal of the “intention to trust.” Cross- and within-brain correlations of this intended-trust signal with neural responses to fiducial markers of the task (investment submitted and investment revealed) identified a remarkable temporal transfer of the “intention to trust” signal from a time just after the revelation of the investor’s decision (a reactive signal) to a time just before this same revelation (an anticipatory signal). This shift suggested that the signal would correlate with the development of a model of the investor in the trustee’s brain. To examine this latter possibility, we ran a separate behavioral experiment ($n=21$ pairs) to test the trustee’s ability to accurately guess (to within $\pm\$1$) the decision by the investor. The error rate of these accurate guesses dropped over the same time period during which the temporal transfer of the future trust signal shifted from reactive to anticipatory. This observation is consistent with the interpretation that the observed signals in the trustee caudate reflect the development of a reputation for their partner.

Lastly, we address an important detail about the amplitude differences between the caudate response to impending increases (black traces, Fig. 39) and impending decreases in trust (red traces, Fig. 39). One explanation, supported by the behavioral data, is that increases in

trust (ΔR) may have a greater effect on their partner's subsequent behavior (ΔI) than decreases in trust. If this were the case, an efficient computational system would devote more computational steps, and hence more energy, to deciding the magnitude of an increase in trust relative to a decrease. In this particular version of the trust game, increases in trust by the trustee were correlated positively with changes in investment on the subsequent round by the investor ($r=0.27$). This was not true for decreases in trust, where there was no such correlation ($r=0.00$). The absence of predictive information associated with a decrease in trust suggests that no analogous energetic investment should be made.

Taken together, these results suggest that the head of the caudate nucleus receives or computes information about (i) the fairness of a social partner's decision and (ii) the intention to repay that decision with trust. In early rounds of the game, the "intention to trust" is evident only after an investment is revealed. With experience, this signal shifts to a time preceding the revelation of the investment. This finding is reminiscent of analogous shifts of reward prediction error signals from reinforcement learning (Berridge 2000; Dayan and Abbott 2001; Dickinson and Balleine 2002) that have recently been identified by fMRI in human caudate and putamen (Pagnoni, Zink et al. 2002; McClure, Berns et al. 2003; O'Doherty, Dayan et al. 2003; O'Doherty, Dayan et al. 2004; Seymour, O'Doherty et al. 2004) and are thought to involve outputs of midbrain dopaminergic systems. These prediction error signals were identified using simple conditioning experiments in which lights predict the future delivery of rewards (e.g., squirt of juice or delivery of monetary return) (Schultz, Dayan et al. 1997; Montague, Hyman et al. 2004). The scheme is simple: An initially neutral light is flashed; it causes no change in dopaminergic activity, but the later (surprising) arrival of juice causes a burst of activity in the dopamine neurons. Repeated pairing of light followed at a consistent time later by juice causes two dramatic changes: (i) The response to juice delivery drops back to baseline and (ii) a burst response occurs just after the light is flashed. This temporal transfer of the burst response to the light is thought to represent the future value predicted by the light. The simplicity of these experiments is somewhat beguiling.

The temporal transfer in the conditioning experiments is directly analogous to the temporal shift that we observe in the trustee brain as they build a model of the investor's response, but framed in the context of a social exchange. In the trustee brain, the analog to the light is the cue for the social partner to invest, and the “social juice” is change in investment. We know that positive changes in investment correlate with subsequent positive changes in repayment; a correlation that grows over the rounds of the task. Early in the exchange, the trustee's intention to increase trust occurs after revelation of the investor's decision to increase investment (Fig. 39A); that is, the increased investment is surprising. The intention to increase repayment therefore follows this revelation. As the game proceeds, this “intention to trust” response transfers to a time before the revelation of the investor decision to increase investment. The only open issue for this speculation is why the signal transferred to this particular time. There are several consistent predictors of the revelation of the investor's decision, but the signal backed up in time to occur just before this. This social prediction error interpretation is provocative and consistent but leaves this important question unanswered. The more general hypothesis is that the dopaminergic system can be used to establish more complex goal states (“rewards”) and make more complex predictions through connections from prefrontal cortex onto midbrain and other subcortical structures (O'Reilly, Braver et al. 1999).

It is possible that similar economic exchange tasks could be used to explore social processing deficits in a variety of neuropsychiatric disorders. These include populations that have faulty or missing capacities for building correct models of others (e.g., schizophrenia or autism spectrum disorders) (Hill and Frith 2003; Lee, Farrow et al. 2004), as well as individuals who misattribute motivations and intentions to others (e.g., borderline personality disorder) (Johnson, Hurley et al. 2003).

III.6.6. Methods

Reciprocity. Investments (I) and repayments (R) were scaled by the amount available to be sent (\$20 for I; three times the amount invested for R). Linear regressions identified significant predictors of change in trust for investors (ΔI_j) and trustees (ΔR_j). Three

predictors of ΔI_j were examined: (i) previous repayment ($R_{j-1} : r = 0.02$), (ii) change in repayment ($\Delta R_{j-1} : r = 0.10$), and (iii) previous trustee reciprocity ($\Delta R_{j-1} - \Delta I_{j-1} : r = 0.31$). Three predictors of ΔR_j were examined: (i) previous investment ($I_j : r = 0.10$), (ii) change in investment ($\Delta I_j : r = 0.26$), and (iii) previous investor reciprocity ($\Delta I_j - \Delta R_{j-1} : r = 0.56$). Thus, reciprocity was a stronger predictor than either amount previously sent (I_j or R_{j-1}) or change in amount previously sent (ΔI_j or ΔR_{j-1}). However, it is noteworthy that reciprocity expressed by the investor ($r = 0.56$) was more strongly related to change in trust than reciprocity expressed by the trustee ($r = 0.26$). This difference is likely accounted for by an asymmetry in the structure of the exchange: In each round, the investor can accumulate money (\$20 endowment) without the cooperation of the trustee, whereas the trustee is wholly dependent on the investor's cooperation. This dependency of the trustee on the investor likely results in greater responsivity by the trustee to changes in investor reciprocity.

III.7. Agency Attribution

Note: The work in this section was mostly done by Damon Tomlin and Amin Kayali, and is included here for completeness as it uses the same data. It was published as "Agent-specific responses in the cingulate cortex during economic exchanges" by Tomlin, D., Kayali, A., King-Casas, B., Anen, C., Camerer, C., Quartz, S. and Montague, P.R. in *Science* **312**(5776):1047-1050 (2006).

III.7.1. Background

Social exchange occurs in species ranging from insects to humans (Hamilton 1964; Hamilton 1964; Trivers 1971). In primates, reciprocal interactions with nonkin occur

repeatedly, thus necessitating the capacity to assign social credit or blame for shared outcomes and to act appropriately according to these assignments (Maynard Smith and Price 1973; Axelrod and Hamilton 1981; Nowak and Sigmund 1992). In humans, reciprocity is a central feature of the collection of psychological mechanisms necessary to support social exchange (Trivers 1971); yet, the underlying neural representations of these mechanisms remain murky. In almost all social exchanges, one must detect and accurately track which social agent (who) gets credit for an outcome. Should credit for an outcome be assigned to one's own actions or those of one's partner? Perhaps such assignments are more a matter of degree—assigning the degree-of-credit to some shared outcome. Understanding such agent-specific mechanisms is important, because the assignment of social agency (Frith and Frith 2001; Vogeley, Bussfeld et al. 2001; Kelley, Macrae et al. 2002; Vogeley and Fink 2003; Ochsner, Knierim et al. 2004; Seger, Stone et al. 2004; Lieberman and Pfeifer 2005) appears to break down in a range of mental illnesses (Frith and Frith 1999; Baron-Cohen and Belmonte 2005; Brune 2005).

Social agency computations are also a prerequisite for generating models of others' mental states. This latter capacity, called theory-of-mind, is highly developed in humans and has been shown to activate a consistent set of brain regions in neuroimaging experiments (Brunet, Sarfati et al. 2000; Gallagher, Happe et al. 2000; Wicker, Perrett et al. 2003; Decety, Jackson et al. 2004). Recent work has complemented these theory-of-mind experiments by using interactive economic games as ecologically realistic models for human exchange (Greene, Sommerville et al. 2001; McCabe, Houser et al. 2001; Rilling, Gutman et al. 2002; Eisenberger, Lieberman et al. 2003; Sanfey, Rilling et al. 2003; de Quervain, Fischbacher et al. 2004; Rilling, Sanfey et al. 2004; Bhatt and Camerer 2005; Delgado, Frank et al. 2005; King-Casas, Tomlin et al. 2005; Singer, Seymour et al. 2006). These experiments have elicited not only brain responses in previously described theory-of-mind networks (Rilling, Gutman et al. 2002; Sanfey, Rilling et al. 2003; Rilling, Sanfey et al. 2004), but also have elicited formerly unreported activations along the cingulate cortex that correlate with the revelation of a social partner's decision (Sanfey, Rilling et al. 2003). Although evoked during an economic exchange with another human, these cingulate

activations did not modulate as a function of the fairness of the exchange, nor did they occur in exchanges with computer partners (Rilling, Sanfey et al. 2004).

This lack of sensitivity to measures of outcome suggests that these responses do not encode some metrical aspect of the trade; instead, they are consistent with the social agency computation described above. We tested this possibility directly on the multi-round trust game.

III.7.2. Cross-Cingulate PCA Analysis

Given the previously reported activations in the anterior and posterior portions of the medial cingulate during a social exchange (Rilling, Sanfey et al. 2004), a detailed analysis of the cingulate cortex in each pair of subjects was performed.

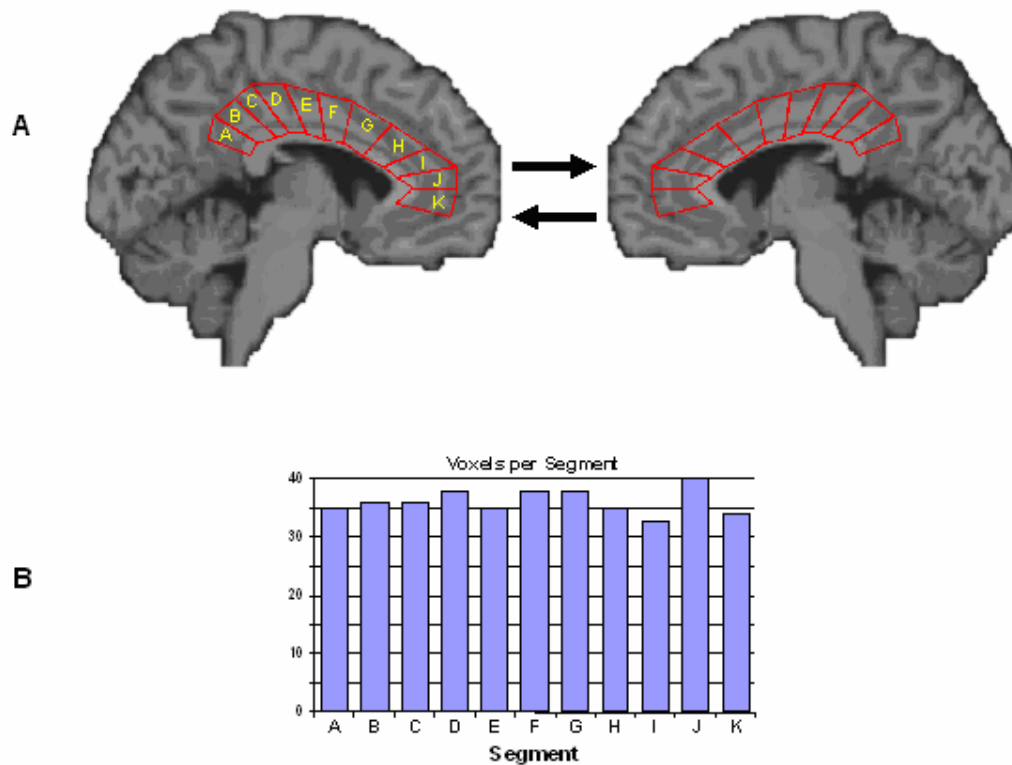


Fig. 41: **Cingulate segmentation.** (A) The population of voxels comprising the cingulate gyrus in both hemispheres was separated into 11 non-overlapping sub-domains based upon their location relative to a predetermined origin. Each group of voxels was designated with a letter, beginning with the most posterior portion

of the gyrus, and rotating through the anterior toward the most ventral part of the cingulate. The boundaries for these zones are shown. The cingulate's width and distance from the origin were not identical for each zone, leading to small variations in the number of voxels comprising each region. The segment boundaries were selected so as to minimize these variations. (B) Total number of functional voxels for each of the 11 cingulate domains examined.

We segmented the medial cingulate and the surrounding paracingulate cortex into separate spatial domains (Fig. 41), computed cross-cingulate and cross-paracingulate correlation matrices for different lags in each phase of the task (investment phase and repayment phase), and carried out temporal principal component analysis (PCA) on the resulting three-dimensional correlation matrix (Fig. 42) (Hyvarinen, Karhunen et al. 2001; Cichocki and Amari 2003). Analysis yielded complementary spatial patterns for cingulate cortices (Fig. 42); that is, patterns of activation in one phase were transposed across role when analysis was performed for the other phase.

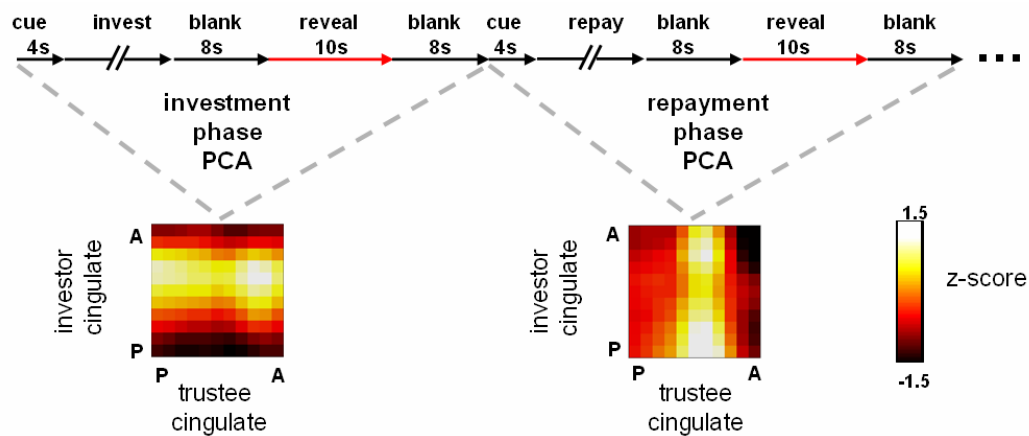


Fig. 42: **Cross-cingulate correlations.** Cross-cingulate principal component analysis (PCA) revealed distinct, but complementary patterns when applied to the cross-correlations between cingulate cortices of investor and trustee.

III.7.3. Differential “Own” and “Other” Responses

The cross-cingulate analysis led us to examine the hemodynamic time series in each cingulate segment. This region-of-interest analysis revealed three distinct response types (Fig. 2A). The first followed the submission of a subject’s own decision (unimodal “own”-

dominated response); the second followed the visual presentation of a partner's decision (unimodal "other"-dominated response). This is a remarkable finding, because visual presentation of the subject's own decision elicited little response in the cingulate cortex. The third response type was bimodal, yielding approximately equal responses after submission of one's own decision and revelation of the partner's decision. However, the peak amplitude of these distinct response types was not uniform across the anterior-posterior axis of the cingulate. Instead, they displayed a systematic spatial variation that was complementary across the basic response types ("own" and "other"). Specifically, the submission of one's own decision elicited maximal activation in middle cingulate regions (Fig. 43A, segment G), whereas viewing the revelation of a partner's decision yielded maximal activation in anterior and posterior cingulate (an example of an anterior response is shown in Fig. 43A, segment K). This result was in stark contrast to the results of the paracingulate analysis, which indicated that, although the dorsal anterior cingulate cortex was highly activated during the experiment, there was no spatial selectivity for either stimulus. In fact, the dorsal anterior cingulate cortex responded strongly to the submission of decisions and the revelation of partner choices, and it was the only paracingulate region significantly activated by either.

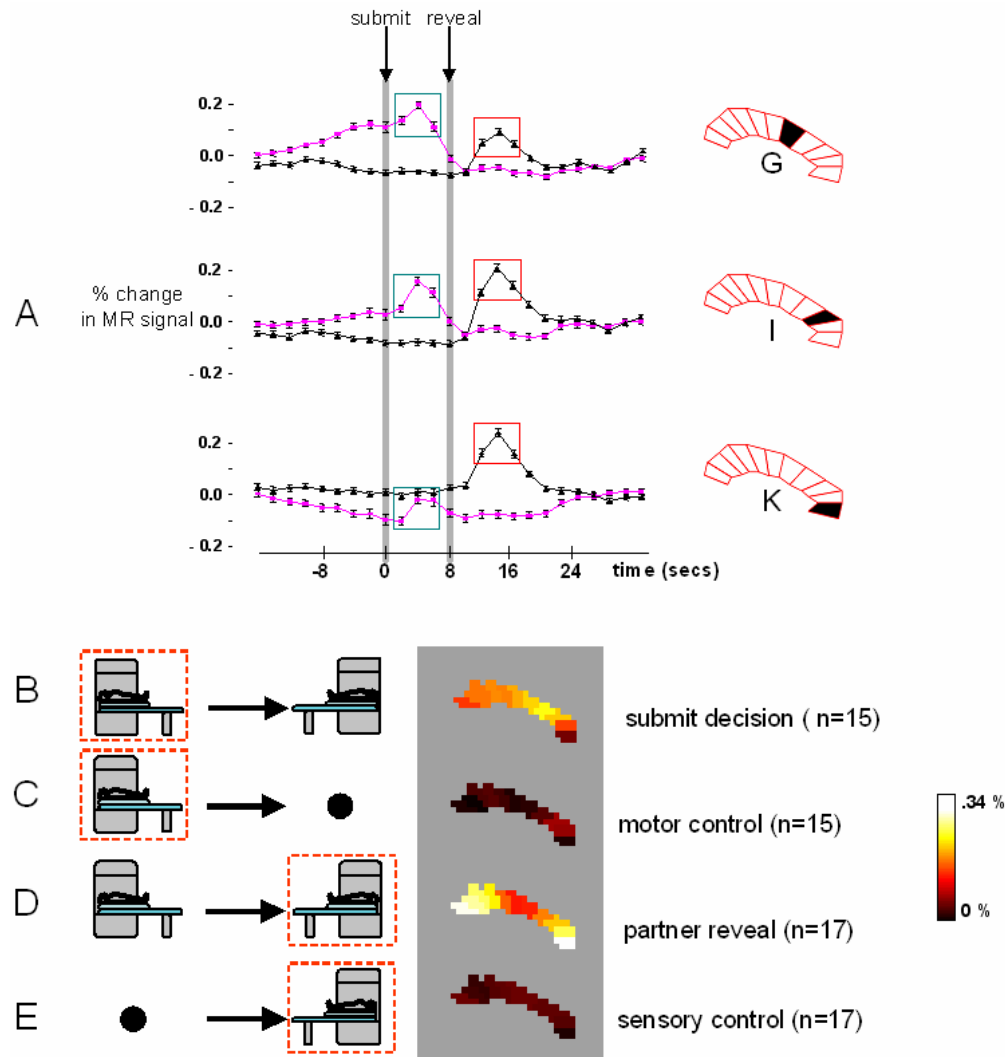


Fig. 43: Agent-specific responses and their pattern disappear outside of economic exchange. (A) Calculation of response pattern diagrams. Traces are the average magnetic resonance (MR) signal during subject decision phases (magenta lines) and during partner decision phases (black lines); error bars represent the standard error of the mean (n=200 subjects). To compute the magnitude of responses to submitting a decision, MR values were selected from the time of peak response and the peak's two flanking points (teal boxes). These values, when averaged, represent the responsiveness of a segment to the submission of the decision. This measure was performed for all segments, and a pseudo color image was produced, as depicted in Figs. 43 and 44. For responses to partner reveal screens, MR values corresponding to the peak activity after screen onset and the peak's two flanking points (red boxes) were averaged and compiled into a similar pseudo color map. (B) The average response to submitting a decision is shown for subjects playing the linked trust game (n=200), and a predominance of the middle cingulate is apparent. (C) Average response profile to submitting decisions in the unlinked motor control experiment (n=15). No significant differentiation was observed across the

cingulate of subjects in this task, but response levels in the middle cingulate were significantly different than those in the linked trust game ($p=0.00001$). (D) Subjects from the linked trust experiment ($n=200$) demonstrate the average response to viewing a social partner's decision. The predominance of responses in the anterior and posterior poles of the cingulate is apparent in this group. (E) Average response to viewing screens in the unlinked visual control experiment ($n=17$). No significant differentiation was observed across cingulate domains, but responses in both anterior and posterior regions were significantly different than those in the linked trust game ($p<0.01$). Maximum activation in (B) and (C) is 0.21% change in MR signal; maximum activation in (D) and (E) is 0.12%; minimum activation for each is 0.00%.

III.7.4. Agent-Specific Responses Disappear in Control Experiments

The distinct response types and the systematic spatial variation of peak amplitudes across the anterior-posterior axis disappeared completely in motor control ($n=15$; Fig. 43C) and sensory control experiments ($n=17$; Fig. 43E) not involving exchange with another agent. In the motor control, subjects reiterated the motor responses of randomly selected investors. We applied the same region-of-interest analysis to the control data (Fig. 43). Statistical comparison of responses in each of the cingulate domains showed that responses differed significantly between the normal trust task and the control tasks. In particular, no significant response was present in the middle cingulate (Fig. 43C), ruling out the possibility that middle cingulate activation in the trust game was the result of motor activity produced by button tapping. In the sensory control, partner reveal screens from the trust game were viewed passively by a separate cohort of subjects ($n=17$). Because partner reveal screens in the trust game had novel content and had been generated by an external agent, we could not use the original data set to separate responses to social or novel stimuli. Thus, subjects in the sensory control task were informed that their compensation depended on money shown under the “gave” label on the screen, but were not told about the social task from which this screen was derived. This manipulation was performed so that a screen's content still held novel and valuable information, but was devoid of social interaction. In each of the 11 cingulate domains, BOLD responses after each of 10 outcome screens did not resemble those obtained during the analogous presentation in the linked experiment (Fig. 43E). There was no systematic spatial variation in response amplitudes across the cingulate gyrus.

III.7.5. Cingulate Pattern Remains Constant for Several Variables

The results provide strong support for three new findings: (i) agent-specific response types localized on the medial bank of cingulate cortex, (ii) a systematic spatial variation of each response type across the anterior-posterior axis of cingulate cortex, and (iii) a dependence of both signals on the presence of a responding agent. Despite the relative simplicity of this economic exchange game, other variable(s) related to this task may have been the underlying cause of the different response types, the spatial variation across the cingulate, and the difference in response to visual revelation of one's own decision and one's partner's decision. However, the different response types and their systematic but complementary spatial variation across the cingulate did not change as a function of a range of dimensions (Fig. 44).

The most dramatic dimensions tested in Fig. 44 are reciprocity and social context (personal versus impersonal). As shown in Section II.6, reciprocity, expressed as degree of tit-for-tat behavior across rounds, acted as a powerful behavioral signal to one's partner and elicited strong, measurable neural correlates. Yet, as illustrated in Fig. 44 (bottom three rows), differences in reciprocity had no effect on the response types or on their spatial variation along the cingulate. The same result held for the difference in social context (personal, $n=104$; impersonal, $n=96$), where prior exposure to one's partner, the sight of their picture in each round, and the knowledge of an imminent encounter afterward had no effect. Likewise, no differences were observed when comparing subject role (investor or trustee), sex of subject, or amount of money sent or received.

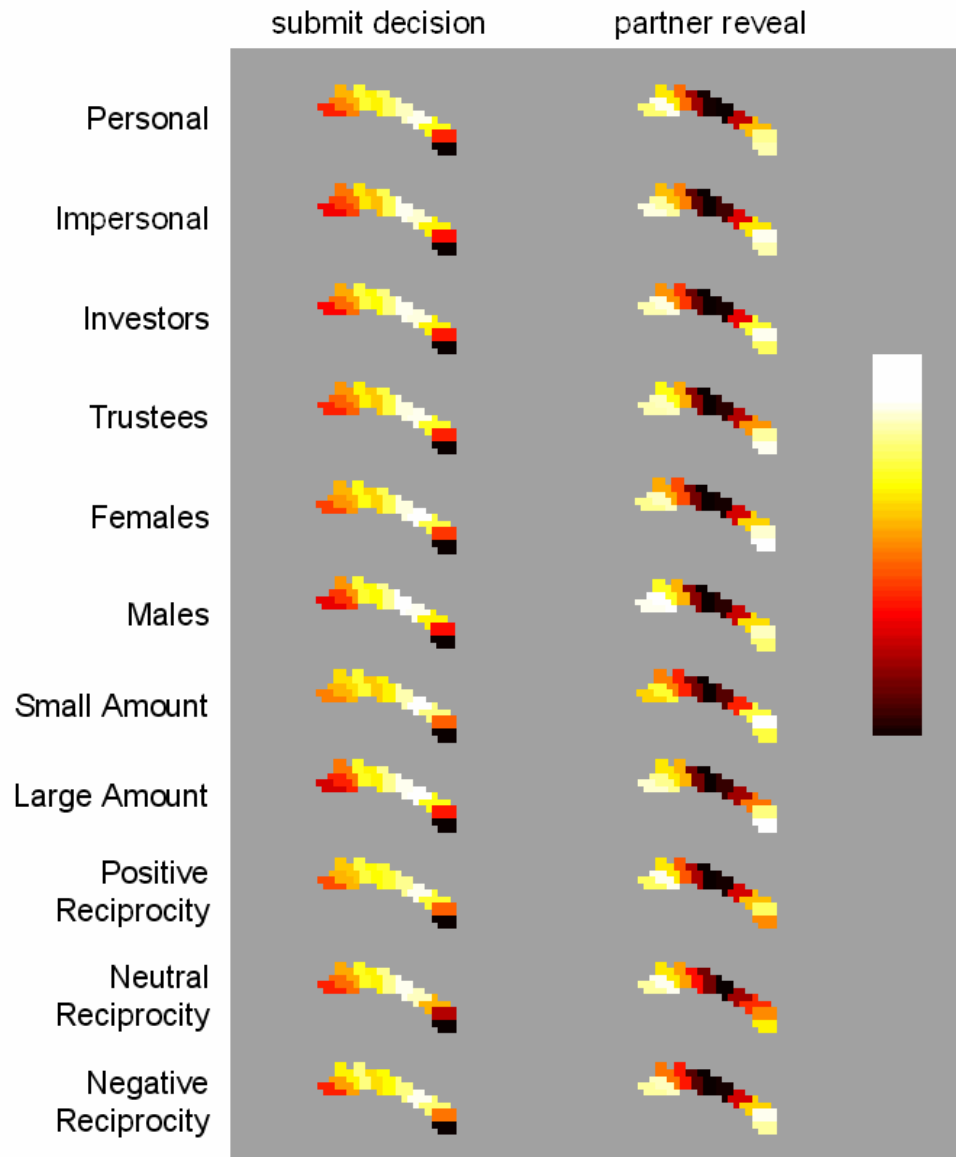


Fig. 44: **Cingulate pattern of “me” and “not me” remains constant across a range of variables.** All responses to decision submission are shown in the left column, whereas those responses to partner reveal screens are shown in the right column. With the exception of the reciprocity and amount diagrams, all responses were averaged across rounds before compilation. Rows labeled “Personal” and “Impersonal” separate activity across social context: the personal ($n=104$ individuals) and impersonal ($n=96$ individuals) tasks. Rows labeled “Investors” and “Trustees” demonstrate the consistency of the responses across the two different roles ($n=100$ for each). Rows labeled “Males” and “Females” demonstrate that these responses do not differ across gender ($n=100$ for each). Rows labeled “Small amount” and “Large amount” show that these patterns do not depend upon the amount of resource sent or received by the player (upper 25% versus lower 25% of payments; $n=454$ and 161 , respectively). Finally, the rows labeled “Positive,”

“Neutral,” and “Negative” reciprocity depict responses across different valences of a behavioral variable of already known interest. These diagrams correspond to average BOLD responses to positive (values > 0.1; n=377 choices), neutral ($-0.1 \leq \text{values} \leq 0.1$; n=865 choices), and negative (values < -0.1; n=458 choices) values of the reciprocity index. Left column maximum is 0.25% change in MR signal; right column maximum is 0.16%; minimum activation for both is 0.00%.

III.7.6. Discussion and Conclusion

Using an iterated economic exchange task, we found two distinct response types along the cingulate cortex consistent with agent-specific responses that signal “me” and “not me”. Rather than residing in strictly demarcated functional zones, these complementary responses types exhibited smooth transitions across the entire medial bank of the cingulate gyrus. It is difficult to probe the extent to which a subject is considering outcomes for oneself or a social partner; individuals in a social exchange must necessarily model the actions of both agents as decisions are made and revealed. Despite this obstacle, the pattern of activation observed in these data was clearly sensitive to which participant was responsible for a given action. The response types and their variation through the tissue space disappeared in control experiments where money sent, actions taken, and money received were matched to those experienced during the normal multi-round exchange (Fig. 43). These controls provide strong evidence that the response types were due to neither motor/premotor responses nor to sensory responses to outcome screens.

One question deserves separate consideration: Did the reveal screens generate simple surprise or novelty responses along cingulate that were not related to the social element of the exchange? Although this reasonable interpretation is possible, the control experiments suggest otherwise. The response pattern along the cingulate disappeared in the control experiments where subjects received stimuli that were visually identical to those in the trust game

and were composed of novel, reward-related information. This manipulation used novel stimuli with economically meaningful content to probe the reveal response and showed neither an “other” response anywhere along the cingulate nor the spatial variation so

prominent in the linked trust task. We take these data as strong support that the responses observed in the linked trust game were not the mere result of surprising content. The response types and their spatial variation along the cingulate were remarkably stable to a range of manipulations. They survived the personal and impersonal treatments, did not change as a function of the reciprocity (see Section III.6.), and were not changed as a function of sex, role, amount sent, or amount received.

The observed lack of change as a function of reciprocity is extremely important because it reduces the likelihood of two alternate interpretations of these data. The average behavior in this game is initial cooperation followed by tit-for-tat moves, a strategy conjectured to be optimal in a reciprocal interaction (Trivers 1971; Nowak and Sigmund 1992). To play such a tit-for-tat strategy, a player's brain must compute the expected next move of their partner and compare this to the actual outcome. Consequently, large deviations in reciprocity would also carry large prediction error signals, a signal type known to show up near or around dorsal anterior cingulate cortex (dACC) (Holroyd and Coles 2002; Holroyd, Nieuwenhuis et al. 2003). Two possibilities arise. The error signals could activate dACC because they reflect directly an error response. Alternately, large deviations in reciprocity represent a signal with a large amount of uncertainty and might engage an output conflict response typical for this brain region (Carter, Braver et al. 1998; Botvinick, Nystrom et al. 1999; Carter, Macdonald et al. 2000; Gehring and Knight 2000; Holroyd and Coles 2002; Critchley, Mathias et al. 2003; Holroyd, Nieuwenhuis et al. 2003; Milham, Banich et al. 2003; Weissman, Giesbrecht et al. 2003). However, neither of these interpretations would anticipate an important feature of the data actually observed. There was no difference in response types or their spatial variation as a function of positive, negative, or neutral reciprocity. One would at least expect both alternate explanations to show responses that differentiated neutral reciprocity from the other two categories (positive and negative). One possibility is that our current analysis missed the error signals altogether for some unidentified reason. However, by using this same behavioral task, we have previously identified such error-related signals elsewhere in the brain and have shown these regions to be sensitive to reciprocity. Consequently, our capacity to detect these error signals

elsewhere makes it less likely that we simply missed error signals in cingulate related to strong deviations in reciprocity. However, it remains a possibility that some unprobed behavioral dimension generated an error signal along cingulate cortex during this task.

In a two-person social exchange, it is crucial for each agent to know how to credit an outcome. Failure to assign this credit accurately will compromise an agent's capacity to decide on an appropriate level of cooperation with the partner—a mistake that could prove extremely costly when averaged over multiple encounters. Consequently, we suspect that these data derive from a neural mechanism dedicated to distinguishing “me” outcomes from “not me” outcomes. The systematic spatial progression of responses suggests to us that this social agency variable may be arrayed as a map; however, the current experiment cannot adequately test this provocative possibility. It is important, therefore, to note that the assignment of credit (or agency) within a social interaction necessarily implicates a variety of cognitive and emotional mechanisms. Thus, although agency parsimoniously characterizes the activations seen with these data, it may not necessarily be congruent to the underlying functions represented along the cingulate.

Extant data support a multifunctional role for the cingulate cortex, particularly in light of the extreme diversity of information that impinges on this region. Cingulate and paracingulate cortices have long been hypothesized as sites of integration of information sources that include cognitive, emotional, and interoceptive signals. Consequently, a range of functions has been ascribed to cingulate cortex (Rainville, Duncan et al. 1997; Carter, Braver et al. 1998; Botvinick, Nystrom et al. 1999; Dougherty, Shin et al. 1999; Bush, Luu et al. 2000; Carter, Macdonald et al. 2000; Damasio, Grabowski et al. 2000; Gehring and Knight 2000; Ochsner, Kosslyn et al. 2001; Critchley, Mathias et al. 2003; Milham, Banich et al. 2003; Phan, Liberzon et al. 2003; Weissman, Giesbrecht et al. 2003; Nielsen, Balslev et al. 2005), and there are disagreements over the exact variables processed and represented in these regions. However, it is reasonably clear that cingulate and paracingulate cortices contribute to normal social cognition and adaptive decision-making (Brunet, Sarfati et al. 2000; Gallagher, Happe et al. 2000; Wicker, Perrett et al. 2003). The results of this paper

add the important possibility that many other variables in the social domain may be arranged in such a systematic fashion through the spatial domain, a phenotype that could be disturbed in afflictions where the capacity to distinguish “me” from “not me” is impaired (Georgieff and Jeannerod 1998; Gallagher, Happe et al. 2000; Johns, Rossell et al. 2001; Lieberman, Jarcho et al. 2004; Ochsner, Knierim et al. 2004; Seger, Stone et al. 2004; Allman, Watson et al. 2005; Baron-Cohen and Belmonte 2005; Brune 2005).

III.8. Conclusions

The feature of having an iterated exchange of money between 2 players makes the multi-round trust game an ideal experiment to study complex interactions and strategies. In this chapter four different aspects have been investigated.

First it was shown that the caudate and the midbrain encode strategic uncertainty, i.e., the ability to predict responses to one’s own behavior. Moreover, the time-courses predicted what strategic choice the Trustee would make later in the round, and the performance of that prediction was assessed on a trial-to-trial basis. Secondly, a novel method was developed to study cross-round and cross-brain interactions, and significant interactions were found in memory-related brain areas (BA9/BA19). Thirdly, the neural correlates of an “intention-to-trust” were found in the caudate, and the peak of those responses shifted forward in time as player reputations developed. Lastly, it was shown that agent-specific responses (“me” vs. “not me”) are arranged in a systematic pattern along the cingulate cortex.

Although these results investigate different aspects of economic decision-making, they answer several important questions (who, how, what?) that allow us to understand the decision-making process as a whole. In order to come up with a justifiable model of human decision-making, it is crucial to investigate and confirm its neuronal and biological validity. This is exactly what these results as well as others in the exciting new field of neuroeconomics try to achieve.

NEURAL CORRELATES OF MORAL DECISION-MAKING

This chapter analyzes various neural aspects of trade-off between efficiency and equity in moral decision-making. It starts off by giving some background about moral decision-making and reviewing the cognitive neuroscience literature on moral judgment. Next the experimental paradigm and setup are described, and the behavioral and neural results are presented. The chapter concludes with a discussion of the main findings.

IV.1. Background

Imagine that you are a doctor who is on duty in a remote rural area. You get a phone call about a patient in location A who is in distress. You promise to help him and drive over to location A. On your way there, you get another phone call about three other people in location B who need help as well. If you continue on your way to location A, you will save the one person, but the other three people will die. Alternatively, if you go to location B, you will save the three people, but the one person will die. Which location would you go to?

This is a typical example of a moral dilemma, where the decision-maker is uncomfortable making either decision. Should I stick to my promise and save the one person, or should I break it and save as many people as possible? Although the definition of morality may vary across cultures, the main definition of moral decision-making includes judgments of rightness or wrongness of actions that cause harm to other people. Two main characteristics typically differentiate a moral dilemma from a non-moral dilemma: (1) the decision-maker is facing a very grave situation (e.g., a life-or-death scenario), and (2) he has to make a

decision between two or more seemingly equal and cruel outcomes. Moreover there is no right or wrong thing to do, there are no unequivocal guidelines, formula or algorithms to follow, and under most circumstances there is no law that tells the decision-maker to favor one outcome over the other. Hence moral decision-making is not a process that is based on a determinate set of rules, but rather on intuition and experience as well as one's own personal moral values.

Most issues in morality and moral decision-making have been around for a long time, and are part of the *big* questions of the *big* thinkers, e.g., Plato or Socrates (who incidentally is often regarded as the father of ethics and moral philosophy in Western societies). Although these questions have been deeply investigated and much work has been published over the centuries, the field has not really advanced and many of the same issues are still alive. Moral decision-making has traditionally been investigated by moral philosophers and cognitive psychologists, who distinguish between several key concepts (e.g., is/ought, act/omit, personal/impersonal), and include emotions and context in order to explain why one action seems morally better than another.

Morality is also a topic of interest in economics, mostly with respect to the distribution of resources (Atkinson 1970; Varian 1975; Wittman 1984; Yaari and Bar-Hillel 1984; Schokkaert and Overlaet 1989; Deiniger and Squire 1998; Konow 2003). Social welfare provision and welfare economics deal with the construction of social welfare functions and are mainly concerned with two issues: economic efficiency ("the size of the pie") and income distribution ("the division of the pie"). Since these issues deal with other-regarding preferences, they can easily be identified as moral dilemmas where a tradeoff between efficiency and distribution needs to be made. These issues are very similar in development economics which deals with aspects of economic growth in developing countries. Again, the tradeoff here is between economic growth of the country and growth of inequality (for example between internal regions of the country). Lastly, moral values also play an important role in public finance, in particular in taxation fairness and efficiency.

More recently, cognitive neuroscientists have investigated moral decision-making and are attempting to solve a long-standing debate in moral philosophy: are moral judgments primarily the result of non-rational and emotional processes or of rational and deliberate reasoning? Recent findings in cognitive neuroscience seem to suggest that both emotion and reason are involved, but that automatic emotional processes (moral intuitions) tend to be stronger.

One of the most famous moral dilemmas that probes moral intuition is the trolley problem (Fig. 45A): A trolley is out of control and is running down a track. On its path there are five people who are working on the track. The only way to save them is to flip a switch that will lead the trolley down a different track on which one person is working. Should you flip the switch and save the lives of five people at the expense of one? Most people agree that it is acceptable to flip the switch. A simple utilitarian calculation justifies this choice, but even non-utilitarians agree that it is acceptable to flip the switch and advance arguments about moral responsibility vs. negligence.

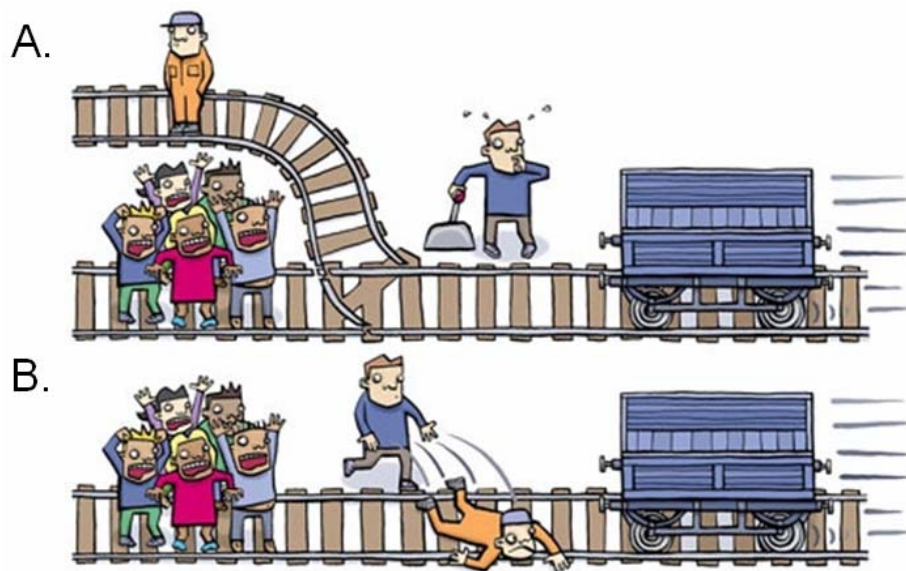


Fig. 45: Classical dilemmas in moral decision-making. A. The trolley dilemma. B. The fat man dilemma (from Big Picture on Thinking, Wellcome Trust, Issue 4, September 2006)

Now consider a very similar scenario, called the fat man dilemma (Fig. 45B): again, a trolley is rolling down a track towards five people. The only way to stop it from crushing the five people is to drop a heavy weight in front of it. As it happens, there is an overweight man standing next to you. If you push him on the tracks he will die, but his body will stop the trolley from crushing the five other people. Should you save five people by sacrificing the life of the fat man? In this scenario, most people say no. This inconsistency creates a puzzle for moral philosophers: how is it morally justifiable to sacrifice one person for five in one scenario but not in the other one? From a utilitarian perspective both scenarios are equal in terms of number of lives that can be sacrificed and saved, but nevertheless people have different moral intuitions. There are several arguments that can be advanced for this inconsistency: in the trolley dilemma there is no direct intention to harm anyone, whereas harming the fat man is part of the plan to save the five people. Another argument is that in the second scenario the fat man is considered to be an innocent bystander, and as such has the right not to be pushed onto the tracks. From a psychological point of view the most convincing argument is that the fat man dilemma engages people's emotions to a greater degree than the trolley dilemma, because it is emotionally more salient to push someone to his death than to flip a switch to produce a similar result.

In one of the first fMRI studies about moral decision-making Greene et al. investigated this hypothesis, and discerned between two different kinds of moral dilemmas (Greene, Sommerville et al. 2001): impersonal moral dilemmas (such as the trolley dilemma, or keeping money found in one's wallet) and personal moral dilemmas (e.g., the fat man dilemma, or stealing a person's organs to help five other people). They found personal dilemmas increased activation in brain areas associated with emotion and social cognition (medial frontal gyrus, posterior cingulate gyrus, and bilateral angular gyrus), and decreased activation in areas associated with working memory (middle frontal gyrus and bilateral parietal lobe) (see Fig. 46). Impersonal dilemmas increased activation in brain areas associated with abstract reasoning and problem solving. When these personal moral dilemmas are further split up into difficult and easy personal moral dilemmas, Greene et al. showed that brain regions associated with abstract reasoning and cognitive control

(dorsolateral prefrontal cortex and anterior/posterior cingulate) are recruited to address the difficult dilemmas (Greene, Nystrom et al. 2004). Moreover several areas in the frontal and parietal cortex were more activated when subjects made utilitarian judgments.

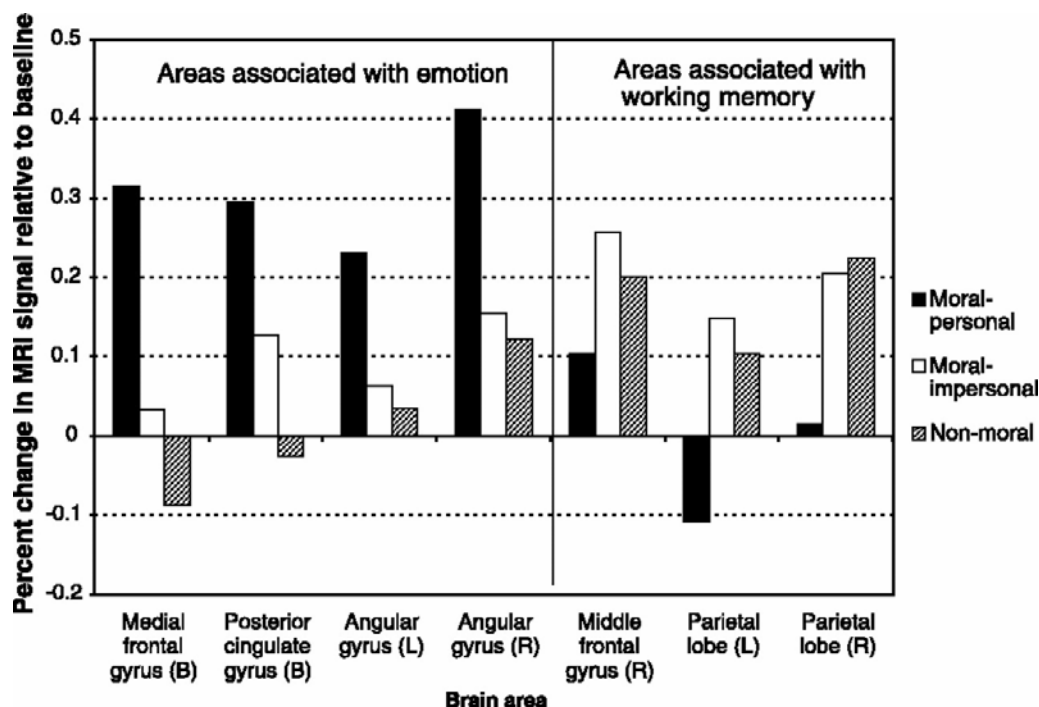


Fig. 46: Activations in personal moral dilemmas. Chart shows brain areas with increased and decreased activation (Greene, Sommerville et al. 2001).

In a 1999 study (Anderson, Bechara et al. 1999) it was shown that patients with early prefrontal cortex lesions demonstrated severely impaired social behavior despite exhibiting basic cognitive abilities. Compared to controls they performed poorly on moral judgment tasks, and largely approached moral dilemmas from the egocentric perspective of avoiding punishment. The importance of the prefrontal cortex is emphasized in another lesion study that shows that patients with lesions in the ventromedial prefrontal cortex produce a utilitarian pattern of judgments on personal moral dilemmas (Koenigs, Young et al. 2007). Several studies have also investigated the neural correlates of moral vs. nonmoral social judgments (Moll, de Oliveira-Souza et al. 2002; Moll, de Oliveira-Souza et al. 2002; Borg, Hynes et al. 2006), and have found that several regions of the prefrontal cortex, as well as

the superior temporal sulcus, are more heavily recruited when considering moral scenarios, providing evidence for the fact that the brain distinguishes between purely emotional and moral situations.

When reviewing the literature in moral decision-making (Greene and Haidt 2002; Greene 2003; Moll, Zahn et al. 2005), it is striking to notice that all judgments were based on hypothetical situations. In each scenario the subject was asked to put himself in an imaginary situation and to decide between two or more imaginary choices. Due to the fact that it is very difficult or even impossible to recreate the gravity of a moral dilemma situation in a laboratory setting, subjects know that their decisions will not have any implications which might compromise their moral judgments as well as their brain activity. After all, there seems to be a difference between what a person thinks he would do in a life-or-death situation and what he would actually do.

The three following criteria help to better understand moral decision-making:

- (i) Real outcomes: Subjects' choices should have a real outcome, i.e., their actions should do some harm (or some good).
- (ii) Parameterization of the decision space: What if I could save 10, 100, 100,000 people by pushing the overweight man onto the tracks?
- (iii) Partition the decision process: Temporally separate the presentation of the dilemma, the subject's decision time, the subject's answer, and the feedback.

We designed a task inspired from the trolley dilemma that satisfies the criteria listed above, and allows the study of various aspects of moral decision-making. The goal of the experiment is to study other-regarding preferences when people have to make trade-offs between different distributional criteria (efficiency and equity).

IV.2. Experimental Design and Methods

IV.2.1. Task

Subjects are asked to make a decision about allocating meals to children from an orphanage in Uganda. There are two types of trials: Give and Take. In Give trials subjects need to decide whether to give k_1 meals to one kid (denoted kid₁) or k_{2a} and k_{2b} meals to 2 kids (denoted kid_{2a} and kid_{2b}). In order to obtain diverse behavior across different types of subjects, the difference between meals is either 0, 1, or 3 meals. In Take trials all children have been endowed with 24 meals, and subjects must decide whether to take away k_1 meals from kid₁ or k_{2a} and k_{2b} meals from kid_{2a} and kid_{2b}. Again, the difference between meals is 0, 1 or 3 meals. In order to vary the level of inequity, the meals are not always split evenly between kid_{2a} and kid_{2b}. Hence the subject must make a decision between the two following choices (also called allocations):

$$[k_1 \quad 0 \quad 0] \text{ or } [0 \quad k_{2a} \quad k_{2b}] \text{ with } k_1 - (k_{2a} + k_{2b}) = \{0,1,3\}.$$

There are a total of 36 trials (18 Give & 18 Take) presented randomly during the experiment. For each group there are 6 trials each with a difference of 0, 1, and 3 meals (see Table 5).

Give Trials			Take Trials		
<i>Kid1</i>	<i>Kid2a</i>	<i>Kid2b</i>	<i>Kid1</i>	<i>Kid2a</i>	<i>Kid2b</i>
23	10	10	-23	-13	-13
23	3	17	-23	-21	-5
23	11	11	-23	-12	-12
23	4	18	-23	-21	-3
23	11	12	-23	-12	-11
23	4	19	-23	-20	-3
19	8	8	-19	-11	-11
19	3	13	-19	-17	-5
19	9	9	-19	-10	-10
19	4	14	-19	-17	-3
19	9	10	-19	-10	-9
19	4	15	-19	-16	-3
15	6	6	-15	-9	-9
15	2	10	-15	-13	-5
15	7	7	-15	-8	-8
15	3	11	-15	-13	-3
15	7	8	-15	-8	-7
15	3	12	-15	-12	-3

Table 5: **Allocation of meals.** The first column denotes the number of meals allocated to the one kid, and columns 2 and 3 denote the number of meals allocated to the 2 kids. Each line represents a different moral dilemma.

IV.2.2. Subjects

24 healthy volunteers (16 female, 8 male) were recruited through Craigslist (www.craigslist.org). Subjects were required to have college education and to be 28–55 years old. Demographics of the subjects are as follows: mean age: 39.2 +/- 5.7 years; marital status: 13 single, 9 married, 2 divorced (4 subjects had kids of their own); education: 16 college, 7 MS, 1 Ph.D. 2 other subjects participated in the experiment, but their data was not used because of motion artifacts in the fMRI data.

IV.2.3. Experimental Setup

The success of this experiment critically depends on the fact that subjects believe that their decisions have a real outcome. To emphasize this, subjects are asked to read through a brochure with a description of the orphanage and a short biography of the children (Fig. 47).


	Bernard xxx	10	Bernard is from southern Uganda. He would like to learn to ride a bicycle and wants a job in medicine when he grows up.
---	-------------	----	---

Fig. 47: **Example of a child's biography** (actual picture and name available upon request)

Next, they are given instructions about the experiment and on how to make their decisions. To give them a sense of their contribution, they are informed that 24 meals correspond to \$5, and that they will be donating around \$60 on average to the orphanage. Throughout the instructions it is stressed that their choices have a real outcome and that meals will be donated according to their decisions.

At the beginning of each trial, subjects are presented with a screen showing whether the trial is a Give or Take trial. They advance to the next screen by pressing a button. Next, they see an animation in which a projectile was moving across the screen toward the kids. The number of meals that each child could potentially receive is denoted next to the picture of the kid. The group that will receive the meals is indicated by the direction of the lever in the middle of the screen. The group of children that get hits by the projectile will receive the number of meals denoted next to the pictures.

When the projectile passes the dotted line, the subject may switch the lever to redirect the projectile towards the other group of kids. He has 3.5 seconds to do so, but can only switch the lever once. After the projectile passes the lever, it continues towards the chosen group of kids. When it hits the box surrounding the kids, the box changes colors and is present for another 3.5 seconds. After a blank screen of random duration (uniformly distributed on 1–3 seconds), a feedback screen of 2 seconds duration informs the subject how many meals each kid received. Trials are separated by a blank screen of random duration (uniformly distributed on 5–7 seconds). Take trials are similar, except that the amounts of meals are negative, and the ball and highlighting is in red instead of green.

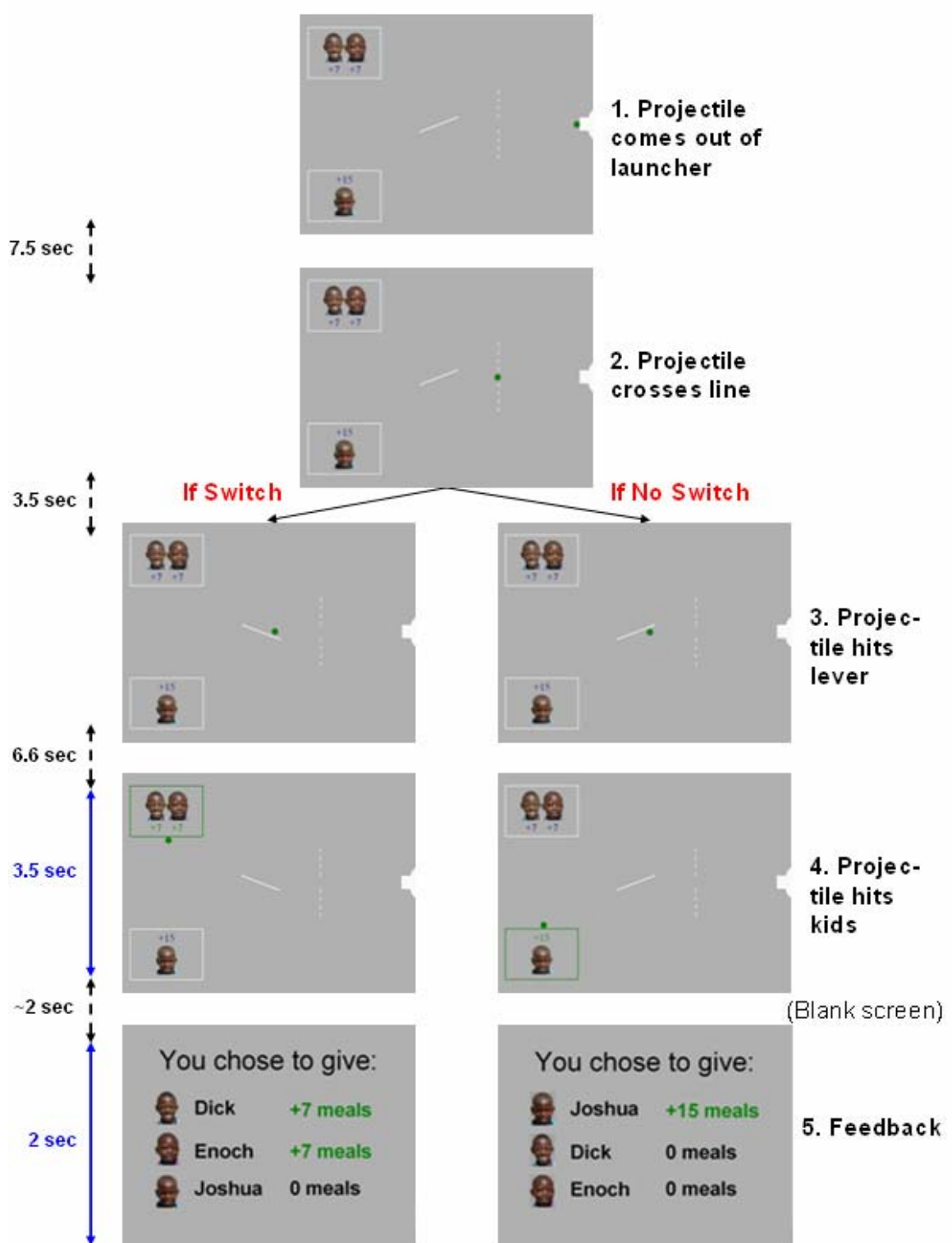


Fig. 48: Timeline of the Moral Decision-Making Task

IV.2.4. fMRI Data Acquisition and Analysis

Brain image acquisition was done on a Siemens Trio. High-resolution T1-weighted scans (1 mm x 1 mm x 1 mm) were acquired using a MPRage sequence. Functional images were acquired using echo-planar T2* images with BOLD (blood oxygenation-level-dependent) contrast, and angled 30 degrees with respect to the AC-PC line. Parameters were as follows: repetition time (TR) = 2000 ms; echo time (TE) = 40 ms; slice thickness = 3 mm yielding in a 64x64x32 matrix (3 mm x 3 mm x 3 mm); flip angle = 90 degs; FOV read = 220 mm; FOV phase = 100 mm, series order: interleaved.

Imaging data was preprocessed using SPM2, and included slice time correction, motion correction, coregistration, normalization to the MNI template, and smoothing of the functional data with an 8 mm kernel (see Section II.4 for details).

GLM analysis was done in SPM2 by specifying a separate general linear model for each subject (fixed effects analysis). First all images were high-pass filtered in the temporal domain (filter width 128 s) and autocorrelation of the hemodynamic responses was modelled as an AR(1) process. In the GLM model all visual stimuli and motor responses were entered as separate regressors that were constructed by convolving a hemodynamic response function (hrf) with a comb of Dirac functions at the onset of each visual stimulus or motor response. The different regressors were at the following moments: when the instruction screen was presented (give or take trial), when the scenario was first displayed, when the subject switched the lever (or for some models when the projectile touched the lever), when the projectile hit the kids and when the feedback screen was displayed. Parametric modulations corresponding to different measures of social welfare (see Section IV.3.2.) were added to the regressors.

IV.3. Behavioral Measures and Behavioral Results

IV.3.1. The Gini Coefficient

The gini coefficient is used to measure the inequality of a distribution (Gini 1912; Gini 1921). Mathematically it is defined as the area between the Lorenz curve of the distribution and the uniform distribution, where the Lorenz curve is the proportion of the distribution assumed by the bottom $x\%$ of the values. For example, for a dataset that includes the income of all households, every point on the Lorenz curve can be described as “the bottom $x\%$ of all households have $y\%$ of all income”. The inequality of income across households is determined by calculating the surface of the area between the uniform distribution and the Lorenz curve (Fig. 49), which is the gini coefficient.

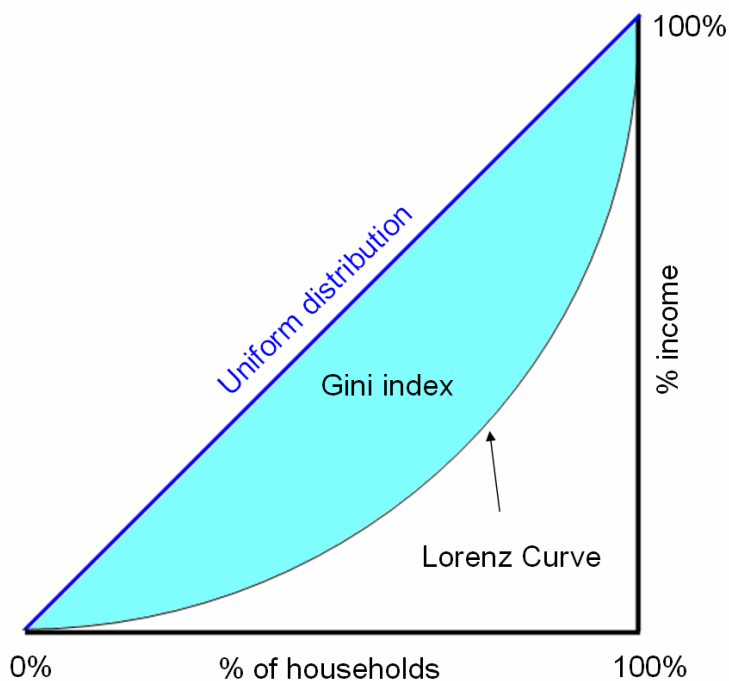


Fig. 49: Graphical representation of the gini coefficient

The gini coefficient varies between 0 and 1, where 0 is perfect equality (uniform distribution), and 1 is perfect inequality (one household has all the income). The gini coefficient is typically used to measure the distribution of wealth in a country (Fig. 50).

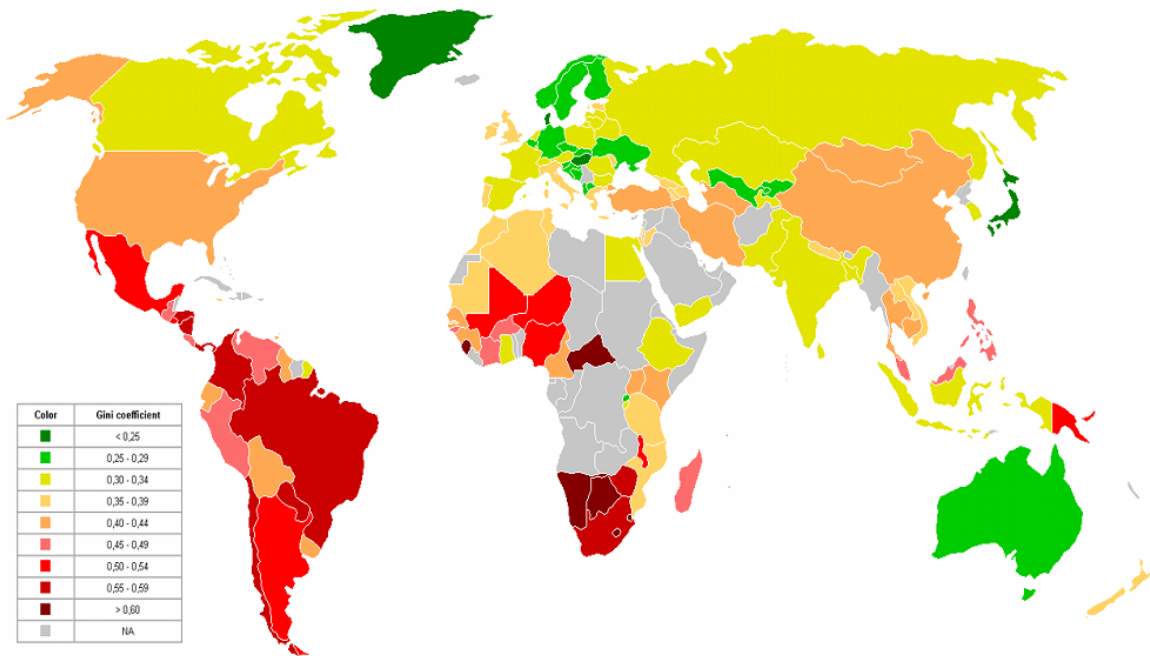


Fig. 50: **Wealth distribution in the World.** Data adapted from wikipedia (http://en.wikipedia.org/wiki/List_of_countries_by_income_equality)

For discrete and unordered data the gini coefficient is calculated as the normalized mean difference between every possible pair of outcomes in the distribution:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n \sum_{i=1}^n x_i}$$

where n is the number of realizations. Compared to other measures of inequality (e.g., the variance), the gini coefficient has the advantage that it is scale independent, e.g., $\text{gini}([2 \ 5 \ 6]) = \text{gini}([4 \ 10 \ 12]) = 0.20$, whereas $\text{var}([2 \ 5 \ 6]) = 4.44$ and $\text{var}([4 \ 10 \ 12]) = 17.33$.

IV.3.2. Measures of Efficiency, Equity, and Utility

The nature of the moral dilemma in the task results from the fact that there are several criteria (number of kids, number of meals and distribution of meals) that cannot all be satisfied simultaneously. The main assumption is that subjects perceive any kind of irregularity in the distribution of meals as unfair, and are thus forced to make a tradeoff between various welfare criteria. In this analysis we distinguish between three different measures of welfare:

1. Efficiency based on the number of meals: subjects try to give the largest number of meals or take away the smallest number of meals. For example, in a scenario where the allocations are [19 0 0] vs. [0 8 8], the subject would give 19 meals the first allocation, whereas he would only give 16 meals in the second allocation. M_c denotes the number of meals in the chosen allocation, and M_u denotes the number of meals in the unchosen allocation.

2. Equity based on the distribution of meals: subjects try to distribute the meals as equally as possible. For example, in the same scenario as above, in the first allocation the first kid receives all 19 meals, whereas the other 2 kids do not receive any. The second allocation is fairer because two kids receive 8 meals each. The distribution of meals is also important within an allocation: for example, the subject's decision might be different depending on whether the second allocation is [0 8 8] or [0 3 13], although both allocations are equal in terms of total number of meals. The distribution of meals within an allocation is measured by the gini coefficient (see previous section for details), and is denoted G_c for the chosen allocation, and G_u for the unchosen allocation.

3. Utility based on a combined measure of the number and distribution of meals: subjects try to find a trade-off between giving the largest number of meals (or taking away the smallest number of meals) and distributing the meals as equally as possible. The utility is calculated as $U = M - \alpha G$ using an inequity aversion model, where α is determined through maximum likelihood estimation from the behavioral data (see Section IV.6. for

methods). Note that the sign on the gini is negative because a large gini coefficient diminishes utility. The utility of the chosen allocation is denoted U_c and the utility of the unchosen allocation U_u .

These three measures can be used to generate three difference measures for both Give and Take trials:

1. Delta efficiency: $\Delta M = M_c - M_u$, denoting the difference of meals between the chosen allocation and the unchosen allocation.

2. Delta equity: $\Delta G = G_c - G_u$, denoting the difference in fairness between the chosen allocation and the unchosen allocation.

3. Delta utility: $\Delta U = \underbrace{(M_c - \alpha G_c)}_{U_c} - \underbrace{(M_u - \alpha G_u)}_{U_u} = \Delta M - \alpha \Delta G$, denoting the difference in utility between the chosen allocation and the unchosen allocation.

Note that all three measures are choice-based, i.e., they compare the levels of welfare of the chosen and unchosen allocations (as opposed to a measure that would for example just measure the difference between the 1 kid and the group of 2 kids, independent of what the subject actually chose).

IV.3.3. Behavioral Data

A summary of the behavioral data from all subjects is shown in Fig. 51. Subjects' choices are partitioned by the absolute difference in meals between allocations (i.e., $|\Delta M|$) and by the nature of the dilemma (Give or Take). The y-axis represents the percentage of subjects who chose to give to or take from kid₁.

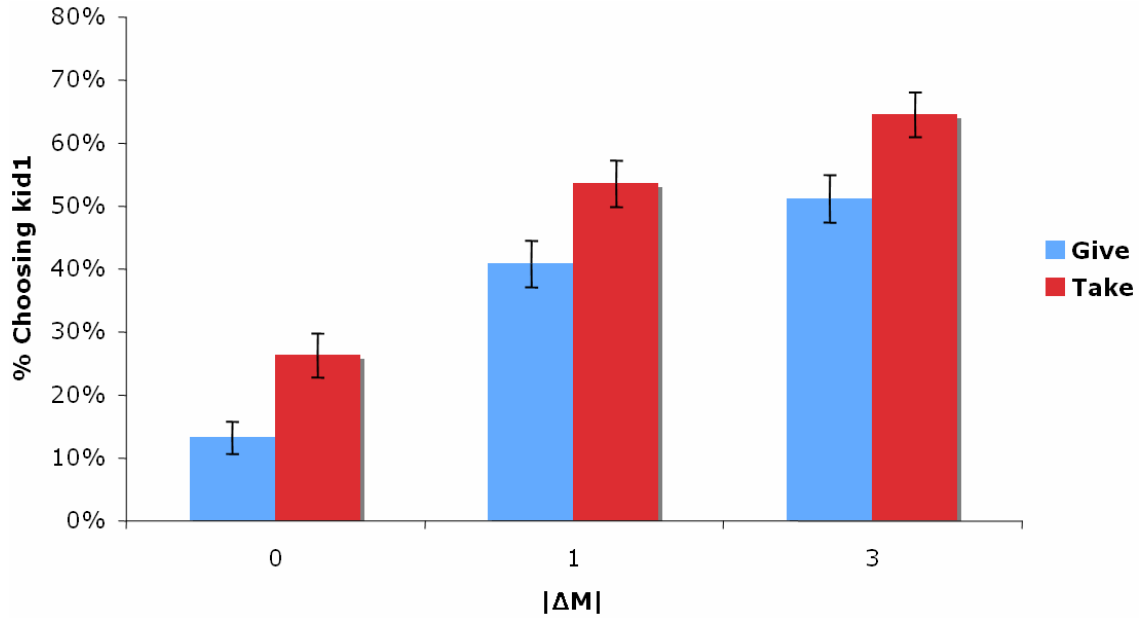


Fig. 51: Subject behavior in the moral task

When the number of meals between two allocations is equal ($|\Delta M| = 0$), only a few subjects choose to give to kid_1 (13% in Give trials) or take meals away from kid_1 (27% in Take trials). They try to help as many kids as possible (helping 2 kids is better than helping 1 kid), and try to avoid hurting 1 kid substantially (hurting 2 kids by a small amount is better than hurting 1 kid by a large amount). Hence they spread the goodness (in Give trials), and dilute the misery (in Take trials).

As the absolute difference in meals increases however, subjects deviate from that rule, and tend to give meals to kid_1 (41% for $|\Delta M| = 1$ and 51% for $|\Delta M| = 3$), and take meals away from kid_1 (53% for $|\Delta M| = 1$ and 65% for $|\Delta M| = 3$). The larger difference in the number of meals between the two allocations leads them to become more utilitarian, i.e. to maximize the total number of given meals, and to minimize the total number of taken meals. It is straightforward to hypothesize that at a sufficiently high value of $|\Delta M|$ all subjects will behave in a utilitarian way, and always choose kid_1 over kid_{2a} and kid_{2b} .

There is also a noticeable difference between Take and Give trials as subjects tend to behave in a more utilitarian way in Take trials. The indifference point ($\sim 50\%$) lies around $|\Delta M| = 3$ meals for Give trials, and $|\Delta M| = 1$ meal for take trials. It should be noted that these results only hold when 15, 19, or 23 meals are given to or taken from kid_1 , and that those indifference points change as a function of the overall number of meals.

IV.3.4. Act/Omit Differences

A well known cognitive bias in psychology is the omission bias: actions are judged to be less morally justifiable than equally harmful omissions (or inactions) (Spranca, Minsk et al. 1991; Baron 1992; Ritov and Baron 1992; Baron 2000). For example withholding the truth seems less immoral than lying, and letting someone die (in cases of euthanasia) seems to be morally more justifiable than killing. There are several good reasons for this, the most compelling one being that an action contains an intention to cause harm, whereas an omission could just be the result of ignorance. Moreover, when omissions are seen as socially acceptable, self-preservation often dominates. For example, under certain circumstances it is acceptable to not jump into water to save a drowning person (whereas it is always immoral to push someone into the river).

This concept of act/omit is important in moral decision-making because the framing of the moral dilemma can cause subjects to switch their decision. For example in the trolley dilemma, some people might argue that they will not switch the lever, because they do not wish to be involved in this difficult decision (they would rather let fate decide). We tested for this possibility of act/omit differences in the moral decision-making task by counterbalancing the location of the kids on the screen as well as the initial direction of the lever to remove any confounding elements.

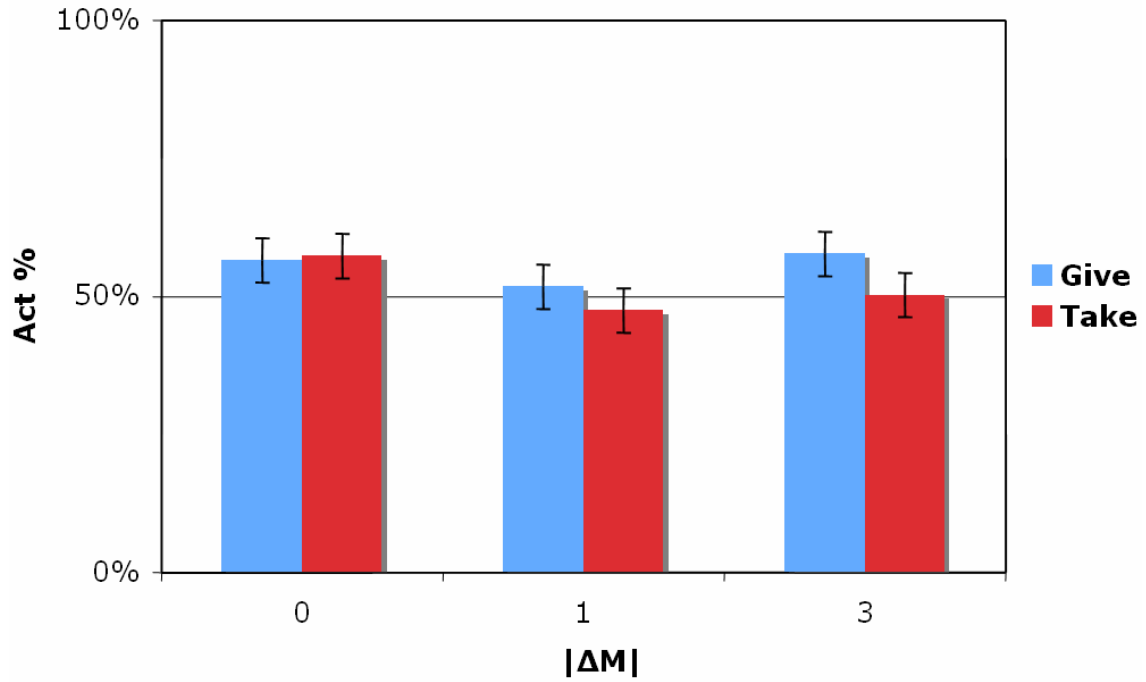


Fig. 52: Act/omit differences in the moral task

As shown in Fig. 52, there are no major differences between action and omission for either Give or Take trials. Subjects acted for about 50% of the trials, and this remains also true when the data is split up according to the absolute difference in meals between allocations $|\Delta M|$. A possible reason for this is that most act/omission differences usually appear when subject behavior is compared across different scenarios. In this task however there is only one scenario type, and the initial direction of the lever might thus not be convincing enough to the subjects to prevent them from acting.

IV.3.5. Inequity Aversion Models

We first estimated the α coefficient from the inequity aversion model by pooling the data over all subjects. This estimate determines the “average” inequity aversion that was found in our subject pool. It was estimated separately for the Give and Take trials: $\alpha_{Give} = 15.27$ and $\alpha_{Take} = 6.96$. These relatively large values of α demonstrate that our subjects were

quite averse to inequity. Furthermore, they were significantly more inequity averse in the Give trials than in the Take trials.

It is clear upon inspection that there is substantial variation within our subjects. Therefore we also estimated the results using choices from individual subjects. The drawback to this method is that, because of the limited number of trials for each subject, there are some subjects whose inequity aversion we were not quite able to robustly estimate. These are subjects who almost always chose to give to one child or take from two children. Therefore we can only place bounds on the level of inequity aversion of those subjects. Nevertheless, it is clear from the estimates that our group results were not driven by a few outliers. Fig. 53 shows the repartition of the individual α coefficients.

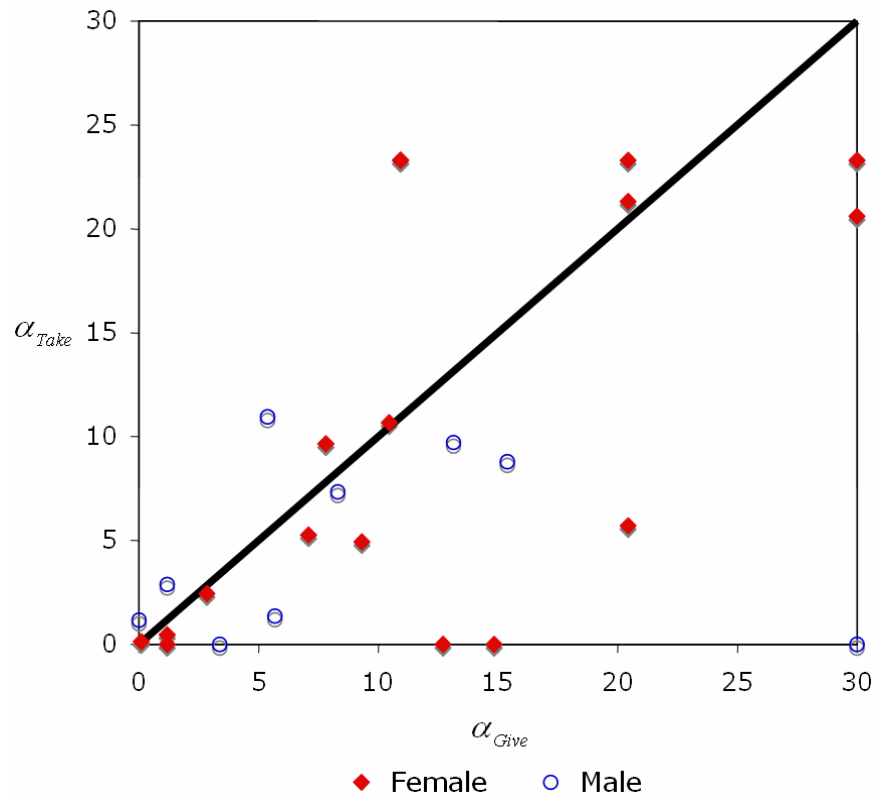


Fig. 53: α coefficient in the inequity aversion model. This data includes 2 subjects that were excluded from the fMRI analysis.

As expected from the pooled analysis, most of the individual data points fall below the diagonal, which confirms that people are more inequity averse in Give dilemmas. This also validates the finding from Fig. 51 that subjects care more about efficiency in Take trials. It can also be seen that female subjects are more equitable in Take trials (higher α_{Take} coefficients) than the male subjects ($p < 0.061$, one-sided). Furthermore, Fig. 54 shows that there is a correlation between the subject's age and α_{Take} ($p < 0.048$), suggesting that older people are more fair-minded in Take trials.

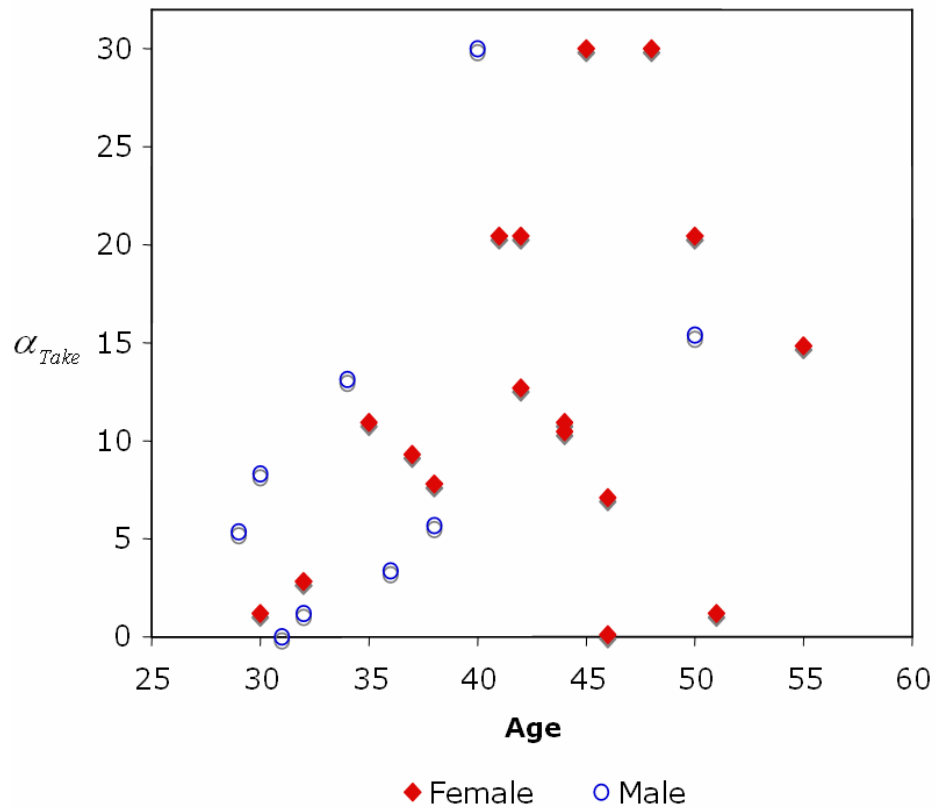


Fig. 54: **Repartition of α_{Take} across age.** This data includes 2 subjects that were excluded from the fMRI analysis.

IV.4. fMRI Results

In this section we present the main neural activations from the Moral Dilemma task.

IV.4.1. *Difference between Allocations*

We looked for brain areas that were correlated with the absolute difference in meals between the two allocations ($|\Delta M|$). Unlike the difference measures in Section IV.3.2., this measure is independent of the subject's choice. $|\Delta M|$ was added as a parametric regressor with 3 values (0, 1, and 3) at the display of the scenario. We found a negative correlation in bilateral insula for both Take and Give scenarios (Fig. 55 and Table 6). The activation in the right hemisphere for the Give trials has only a very small overlap with the insula.

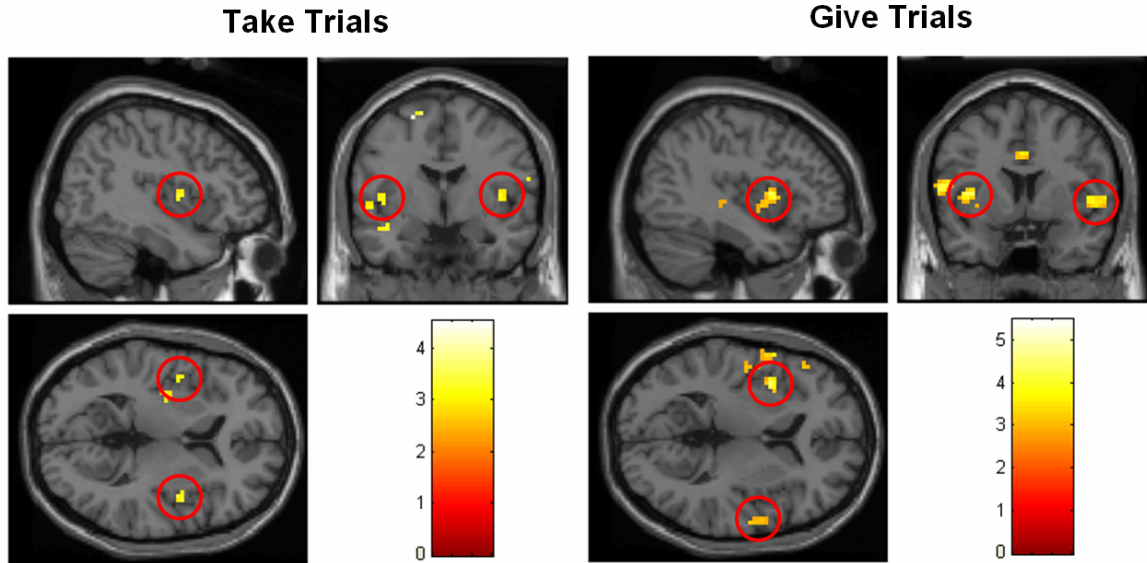


Fig. 55: **Difference between allocations.** Negative parametric activation in bilateral insula for both Take and Give trials with respect to $|\Delta M|$ when the scenario is displayed to subjects. A statistical map is shown alongside a pseudo-color legend with t-scores ($p \leq 0.005$, minimum cluster size: 5).

Scenario	Area	X	Y	Z	T	#Voxels
Take	Left Insula	-42	0	3	3.41	33
	Right Insula	42	-3	6	3.26	15
Give	Left Insula	-39	6	6	4.40	42
	Right Insula	48	6	3	3.65	38

Table 6: **Activations in the insula for $|\Delta M|$.** X, Y, Z = MNI location coordinates of peak voxel (mm); T = T-statistic of peak voxel; #Voxels = number of activated voxels in the cluster ($p \leq 0.005$, minimum cluster size: 5)

Since $|\Delta M|$ is negatively correlated with the brain activity in the insula, this implies that the larger the difference in meals, the smaller the activation in the insula. To verify this, we extracted the time-courses from the activated voxels in the insula, and segregated them according to the difference in meals between allocations. At about 6 seconds after the onset of the display screen the time-courses separate, indicating the insula is differentially activated with respect to $|\Delta M|$. Fig. 56 shows this for Take trials (the time-courses are similar for Give trials, although more noisy).

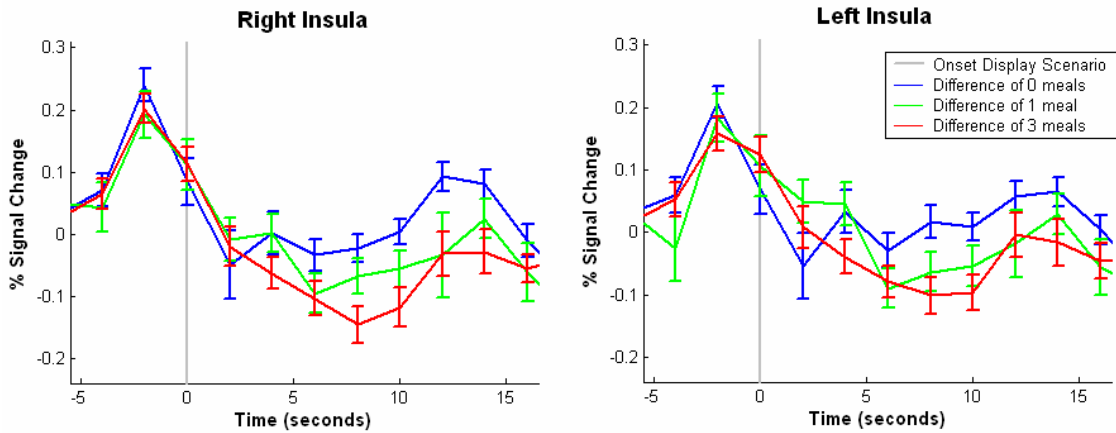


Fig. 56: Time-courses in bilateral insula in Take trials

IV.4.2. Correlates of Efficiency, Equity and Utility in Take Trials

We next investigated whether there are any neural correlates of the welfare measures described in Section IV.3.2 in the Take scenarios. We were not able to use those measures directly as there is a correlation between the chosen and unchosen allocation. Instead, we used the difference measures (i.e., ΔM , ΔG , and ΔU), which measure the actual spread in efficiency, equity, and utility between the chosen and unchosen allocations. We hypothesize that various brain structures might track these measures.

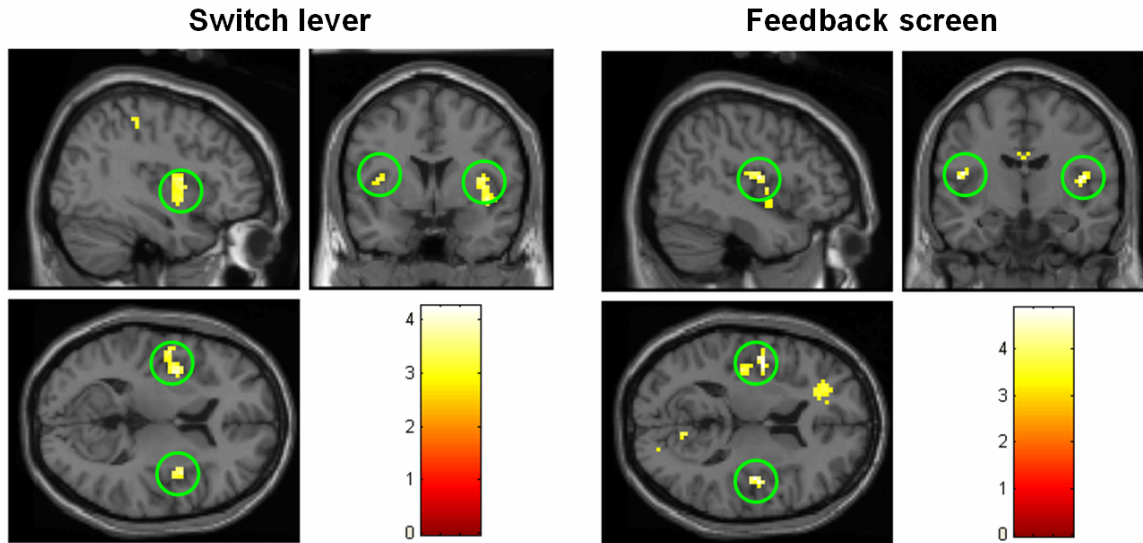


Fig. 57: **Delta equity in Take trials.** Negative parametric activation in bilateral insula with respect to ΔG when subjects switch the lever and when the feedback screen is displayed. A statistical map is shown alongside a pseudo-color legend with t-scores ($p \leq 0.005$ for switch lever, and $p \leq 0.002$ for feedback, minimum cluster size: 5).

Screen	Area	X	Y	Z	T	#Voxels
Switch Lever	Left Insula	-36	-3	12	4.16	27
	Right Insula	39	3	9	3.72	64
Feedback	Left Insula	-45	-6	12	4.89	18
	Right Insula	45	-9	12	4.74	26

Table 7: **Activations in the insula for ΔG during Take trials.** X, Y, X = MNI location coordinates of peak voxel (mm); T = T-statistic of peak voxel; #Voxels = number of activated voxels ($p \leq 0.005$ for switch lever, and $p \leq 0.002$ for feedback, minimum cluster size: 5). Note that only the voxels that actually are in the insula were considered in #Voxels.

We used the difference measures to construct parametric regressors at various moments in the experiment, and looked for brain areas whose activation was correlated with those regressors. We found that the bilateral insula was negatively correlated with delta equity (ΔG) when subjects switched the lever, as well as when the feedback screen was presented (Fig. 57 & Table 7). There was also a weak correlation in the right insula when the projectile hit the lever. We also found that delta utility (ΔU) was positively correlated with brain activity in the caudate at the moment where the projectile hits the kids (Fig. 58).

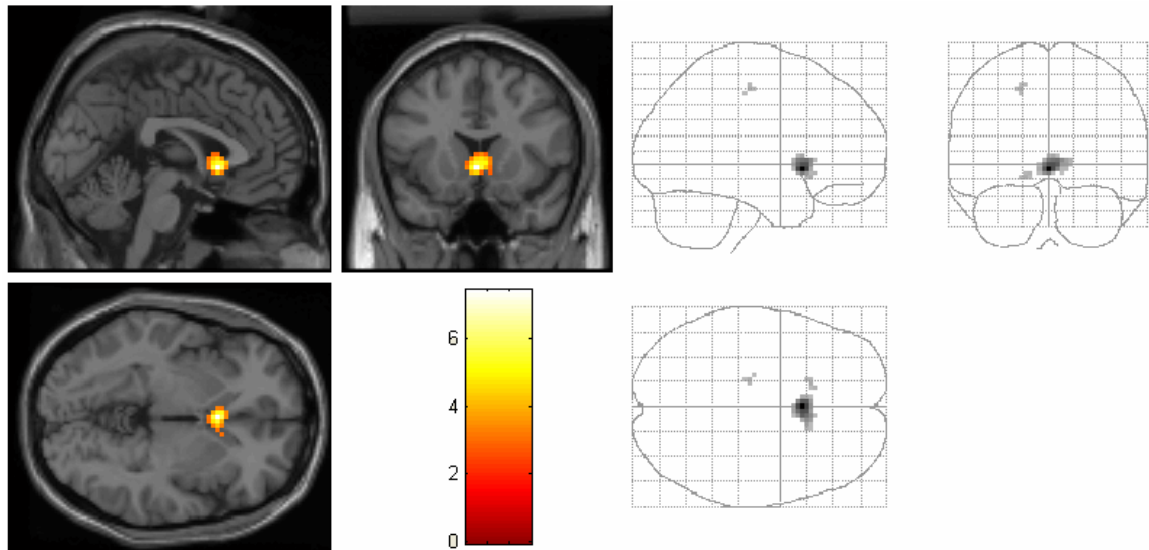


Fig. 58: **Delta utility in Take trials.** Positive parametric activation in the caudate with respect to ΔU when the projectile hits the kids. The left figure panel shows a statistical map alongside a pseudo-color legend with t-scores ($p \leq 0.005$, minimum cluster size: 5). The right panel is a glass brain of the same activation, showing that the only activated brain structure is the caudate (84 voxels).

IV.4.3. Correlates of Efficiency, Equity and Utility in Give Trials

We also looked for similar activations with respect to the difference measures in Give trials, and found a positive correlation between delta equity (ΔG) and brain activity in the bilateral caudate. This correlation was present at several instances during scenarios: when the scenario was first displayed, when the subject switched the lever, when the projectile hit

the lever, and when the projectile hit the kids (see Table 8 for a summary). Although the peaks of activation were slightly different for each instance, there was quite a lot of overlap between voxels. Activation in the caudate when the projectile hits the kids is shown in Fig. 59.

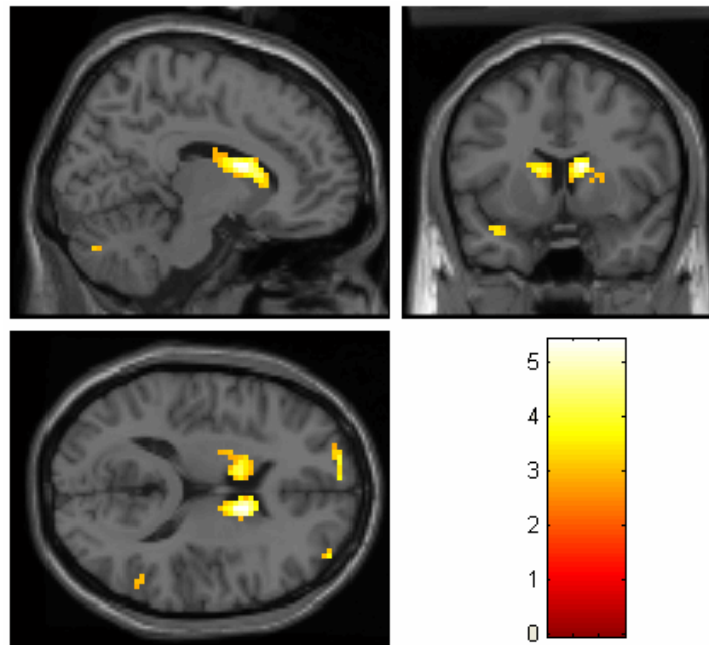


Fig. 59: **Delta equity in Give trials.** Positive parametric activation in the caudate with respect to ΔG when the projectile hits the kids. Statistical map shown alongside a pseudo-color legend with t-scores ($p \leq 0.005$, minimum cluster size: 5)

Screen	Area	X	Y	Z	T	#Voxels
Display scenario	Right caudate	18	18	12	3.74	37
Switch lever	Left caudate	-6	12	9	3.25	9
	Right caudate	9	15	6	4.67	19
Hit lever	Left caudate	-3	6	6	4.14	48
	Right caudate	6	6	6	4.48	35
Hit kids	Left caudate	-12	6	12	4.27	39
	Right caudate	12	9	15	5.41	85

Table 8: **Activations in the caudate for ΔG during Give trials.** X, Y, X = MNI location coordinates of peak voxel (mm); T = T-statistic of peak voxel; #Voxels = number of activated voxels ($p \leq 0.005$, minimum cluster size: 5). Note that only the voxels that actually are in the caudate were considered in #Voxels.

Parametric correlations with respect to delta efficiency (ΔM) were found in the left thalamus (MNI coordinates of the peak: [-9 -9 9], 40 voxels, $p < 0.001$) when the feedback screen was presented to subjects (Fig. 60), a region implicated in the evaluation of reward.

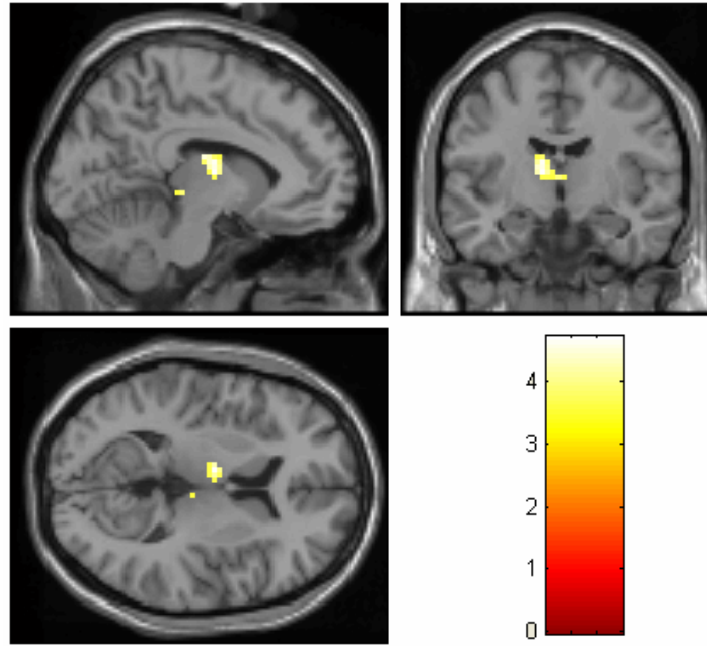


Fig. 60: **Delta efficiency in Give trials.** Positive parametric activation in the caudate with respect to ΔM during the feedback screen. Statistical map shown alongside a pseudo-color legend with t-scores ($p \leq 0.001$, minimum cluster size: 5)

IV.5. Discussion and Conclusion

In this study we standardized and parameterized various criteria of welfare (efficiency, equity, and utility) in the context of a moral decision-making task, and correlated the behavioral welfare measures with neural activity. There were three brain areas in particular that showed significant correlations:

- (i) Insula: absolute delta efficiency ($|\Delta M|$), delta equity (ΔG) in Take trials
- (ii) Caudate: delta utility (ΔU) in Take trials, delta equity (ΔG) in Give trials
- (iii) Thalamus: delta efficiency (ΔM) in Give trials

Several of those areas were activated at different moments during the game, but for the purpose of this study we do not intend to make any claims about the temporal evolution of the decision-making process.

IV.5.1. Insula Activations

The insula is part of the limbic system, which includes the brain structures that are involved in emotion, motivation and emotion-related memories. It receives input from various cortical areas and the thalamus, and it delivers information to sensory areas, the orbitofrontal and cingulate cortices, as well as to limbic structures such as the amygdala, the ventral striatum or the hypothalamus. As such the insula is involved in linking emotional and sensory information, and has been hypothesized to represent internal bodily states (Damasio, Grabowski et al. 2000). More specifically, fMRI experiments have shown that the insula is activated under several negative emotional states such as pain, fear, disgust, anger or sadness.

We first showed that insula activation was negatively correlated with the absolute delta efficiency ($|\Delta M|$) (Fig. 55) at the moment when the moral dilemmas are first revealed to the subjects for both Give and Take trials. This measure is choice independent, suggesting that the insula is merely evaluating the spread in meals between dilemmas, and differentiating between fair offers (offers that give the same number of meals to kid₁ and kid_{2a} and kid_{2b}) and unfair offers (offers that give an equal number of meals to kid₁ and kid₂). A similar activation was found in the Ultimatum game (Sanfey, Rilling et al. 2003), where unfair offers elicited increased activity in the insula. Although the task did not have any moral components, the concepts of fairness are very similar in both cases.

We also found that the insula was parametrically correlated with delta equity (ΔG) in Take trials at various instances throughout the experiment (when the subject switches the lever, when the ball hits the kids, and during the feedback screen) (Fig. 57). The gini coefficient is a measure of inequality, and delta equity thus measures the spread in inequality between the chosen and unchosen allocations. A positive delta equity coefficient means that the chosen allocation is more *unequal* than the unchosen one, and a negative delta equity

means that the chosen allocation is more *equal* than the unchosen one. Activity in the insula is negatively correlated with the delta equity coefficient, meaning that the insula is more activated when the subject picks the more equal allocation ($\Delta G < 0$), and less activated when he picks the more unequal allocation ($\Delta G > 0$).

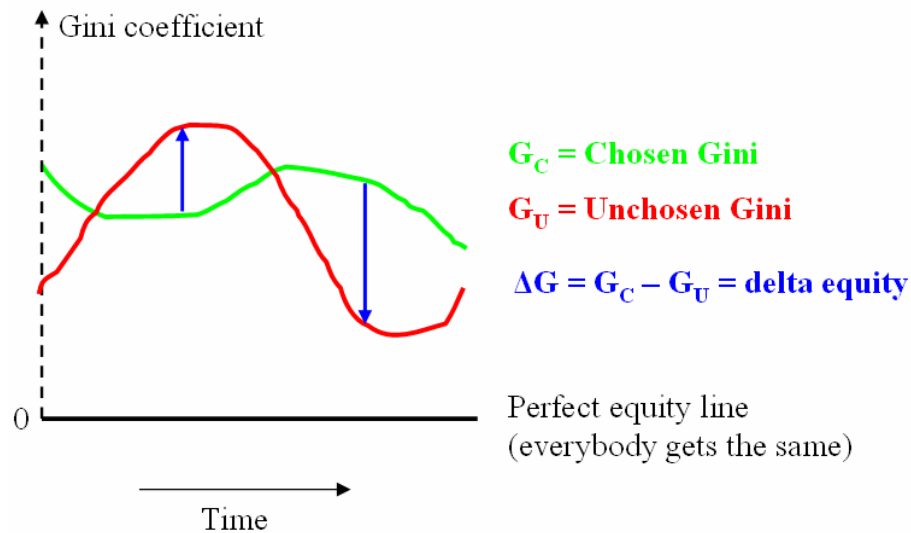


Fig. 61: **Interpretation of delta equity.** Note that the G_C and G_U are always above the perfect equity line, i.e., a distribution cannot be more equal than perfect equity.

Generally speaking, this activation can be interpreted as a deviation from a norm—in this case the norm is equality, i.e., sharing the resources equally between all recipients (Fig. 61). In the forced choice scenario from the experiment this is not an option though, since the subject cannot choose to distribute the resources equally. He has to decide between two allocations that do not have an equal level of equity. Hence the insula measures the spread in fairness between the two allocations according to the subject's choice.

At first glance it might seem paradoxical that activation in the insula is negatively correlated with delta equity, because the insula is generally activated under negative conditions. It should thus be more activated when the subject picks the more unequal allocation which is the opposite of what we observe. But our finding makes sense if we consider that the insula is not just passively coding for differences in the inequality, but is

actively participating in the decision. That is, high insula activity could “cause” the subject to choose the low inequality allocation. Alternatively, low insula activity allows the subject to choose the high inequality allocation. Support for this explanation comes from the same study about the Ultimatum game (Sanfey, Rilling et al. 2003) where insula activation was found to be positively correlated with the rejection rate of unfair offers. This was interpreted as the “insula causing rejection”. In the Ultimatum game, the responder can either reject or accept an unfair offer, resulting in allocations of $(0,0)$ and (x_1, x_2) respectively (where $x_1 \neq x_2$). If we look at this in terms of equity we get $Gini(0,0) < Gini(x_1, x_2)$, or, equivalently, $Gini(reject) < Gini(accept)$, resulting in a negative delta gini if the responder rejects the offer, and in a positive delta equity if he accepts the offer. Since rejection is associated with increased activity in the insula, this is precisely the same as what our results demonstrate.

IV.5.2. Caudate Activations

We found a positive parametric correlation between activity in the bilateral caudate and delta utility (ΔU) when the projectile hits the kids during Take trials (Fig. 58). Although the caudate is thought to be mainly involved with the control of voluntary movement, it has also been shown to play an important role in the brain’s learning system. More specifically, reward prediction errors from reinforcement learning have been identified in the human caudate and are thought to involve outputs of the midbrain dopaminergic systems (McClure, Berns et al. 2003; O’Doherty, Dayan et al. 2003; Seymour, O’Doherty et al. 2004; Knutson and Cooper 2005; Haruno and Kawato 2006). Hence the activation based on delta utility can be interpreted as an evaluation of the differential reward associated with the two allocations. Although subjects make their decision earlier in the scenario, the actual outcome is revealed at the moment when the projectile hits the kids. There is no uncertainty associated with this outcome, so there is no reward or utility updating per se, but the caudate would still be evaluating the difference in utility between the actual outcome and the hypothetical outcome if the choice had been reversed.

We also found a very consistent positive parametric activation with respect to delta equity (ΔG) in the bilateral caudate during Give trials (Fig. 59). Since a positive value of delta equity means that the subject chose the more unequal allocation, this activation cannot be interpreted in terms of reward. However if we look at it in terms of punishment, the sign of the activation makes more sense. Several studies have found the caudate to be activated under punishment conditions (Seymour, O'Doherty et al. 2004; Seymour, O'Doherty et al. 2005). The caudate has also been shown to have a larger response in a punishment condition than in a reward condition early at the onset (Delgado, Nystrom et al. 2000; Delgado, Locke et al. 2003). Although looking at punishments rather than rewards might seem arbitrary, it is justified by the nature of the moral dilemma, which makes either choice feel like a punishment.

IV.5.3. Conclusions

In this study we have found neural correlates for various measurements of welfare (delta equity, delta efficiency, and delta utility) in a moral decision-making task. The power of this study comes from two design characteristics: (i) subjects' decisions have real outcomes as opposed to the hypothetical scenarios usually used in moral decision-making studies; and (ii) the variables associated with equity and efficiency have been parameterized to allow for a parametric analysis of the neural data. We found evidence for the direct involvement of these essential measures of welfare in the decision-making process. In particular, activity in the caudate was directly correlated with a differential measure comparing the two allocations. Moreover, insula activation was shown to lead subjects to choose the more equitable of two allocations. This analysis suggests that although moral decision-making seems to be more difficult than other types of decision-making, the brain might treat it in a very similar way.

IV.6. Methods

Inequity Aversion Model

A simple inequity aversion model was used to assess the tradeoff between equity and efficiency. Subjects' utility functions were assumed to be $u(\mathbf{x}) = \sum_{i \in I} x_i - \alpha \cdot \text{gini}(\mathbf{x})$, where \mathbf{x} is a vector of allocations for the kids. Therefore, efficiency is represented by the total number of meals of the allocation, and inequality by the gini function.

The utility function makes several strong assumptions about the tradeoff between equity and efficiency. First, it assumes that people value efficiency linearly, whereas typically some diminishing marginal utility is observed. Because the range of meals used is rather small (see Table 5), diminishing marginal utility is not a parsimonious explanation. Second, we assumed that people value the sum rather than some measure of centrality, such as the mean. In our choices subjects always chose allocations over 3 people, and hence it makes no difference whether choosing the mean or the sum.

The probability that the subject chooses allocation \mathbf{x}_1 is given by (according to the logit or softmax formula):

$$P(\mathbf{x}_1, \mathbf{x}_2; \alpha, \lambda) = \{1 + \exp(-\lambda(u(\mathbf{x}_1; \alpha) - u(\mathbf{x}_2; \alpha)))\}^{-1}$$

The parameter λ is the sensitivity of choice probability to the utility difference (the degree of inflection), or the amount of “randomness” in the subject's choices ($\lambda = 0$ means choices are random; as λ increases the function is more steeply inflected at zero).

Denote the choice of the subject in trial i by y_i , where $y_i = 1$ if subject chooses allocation \mathbf{x}_1 , and 0 otherwise. We fit the data using maximum likelihood, with the log likelihood function:

$$\sum_{i=1}^N y_i \log(P(\mathbf{x}_1, \mathbf{x}_2; \alpha, \lambda)) + (1 - y_i) \log(1 - P(\mathbf{x}_1, \mathbf{x}_2; \alpha, \lambda)).$$

A Nelder-Mead simplex algorithm (Nelder and Mead 1965), implemented in Mathematica v5.2, was used to find the maximum. Ten random starting positions were used and the iteration with the highest likelihood value was chosen.

CONCLUSION

One of the major goals of neuroeconomics is to test the biological validity of existing economic frameworks and to use the neural data to create more accurate models of human behavior. As such neuroeconomics and neuroscience are closely interconnected and can be mutually beneficial. The brain can be seen as a black box that computes a set of outputs for a set of inputs. However, the brain is not deterministic, i.e., it is possible that the brain computes a different set of outputs for the same set of relevant inputs. This means that there is some randomness associated with our decisions. Moreover, brains from different people also function differently. Behavioral economics has been trying to figure out how the black box works by choosing well-determined sets of inputs and studying the corresponding outputs. The technological advances in neuroscientific tools allow us to probe very specific parts of the black box, and to understand how those outputs are formed. But human behavior is often complex, unpredictable, and inconsistent, and it is thus necessary to study it under a multitude of well-defined rules and outcomes.

This thesis investigated two such paradigms, namely the neural correlates of cooperation in a two-person dynamic game, and the neural correlates of moral decision-making. In Chapter 2 we showed that economic concepts such as reciprocity and strategic uncertainty have neural correlates, and how social variables such as trust and agency develop in the brains of two interacting players. In Chapter 3 we showed how various brain structures encode measures of efficiency, equity and utility in a moral dilemma situation. In order to understand a structure as complex as the human brain it is also necessary to use a variety of methods. In Chapter 2 we developed alternative methods to analyze fMRI data, and more specifically to analyze synchronized fMRI data from two interacting brains.

Although the results presented in this thesis have only been verified in the context of the two specific experimental set-ups, similar studies have already or will confirm the results under different experimental conditions. It is the combination of results from multiple studies that will eventually allow us to understand how people build trust, design strategies or evaluate different options in order to make a decision. Although the current key objective of neuroeconomics is to understand how the brain works, the neuroeconomic findings will inevitably also have other uses, and in particular clinical applications. For example if we understand the neural correlates of trust, it might be possible to help people who have trouble maintaining interpersonal relationships (e.g., sociopaths). Although it is doubtful that we will ever fully understand how the brain works (is a system capable of understanding itself?), constantly pushing our knowledge to its limits should remain the ultimate goal.

BIBLIOGRAPHY

- Adolphs, R. (2003). "Cognitive Neuroscience of Human Social Behaviour." *Nat Rev Neurosci* **4**(3): 165-178.
- Allman, J. M., Watson, K. K., Tetreault, N. A., and Hakeem, A. Y. (2005). "Intuition and Autism: A Possible Role for Von Economo Neurons." *Trends Cogn Sci* **9**(8): 367-373.
- Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1999). "Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex." *Nat Neurosci* **2**(11): 1032-1037.
- Andreasen, N. C., O'Leary, D. S., Arndt, S., Cizadlo, T., Hurtig, R., Rezai, K., Watkins, G. L., Ponto, L. L., and Hichwa, R. D. (1995). "Short-Term and Long-Term Verbal Memory: A Positron Emission Tomography Study." *Proc Natl Acad Sci U S A* **92**(11): 5111-5115.
- Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., and Poldrack, R. A. (2004). "Human Midbrain Sensitivity to Cognitive Feedback and Uncertainty During Classification Learning." *J Neurophysiol* **92**(2): 1144-1152.
- Ashburner, J., and Friston, K. J. (1999). "Nonlinear Spatial Normalization Using Basis Functions." *Hum Brain Mapp* **7**(4): 254-266.
- Atkinson, A. (1970). "On the Measurement of Inequality." *Journal of Economic Theory* **2**: 244-263.
- Axelrod, R., and Hamilton, W. D. (1981). "The Evolution of Cooperation." *Science* **211**(4489): 1390-1396.
- Baron-Cohen, S., and Belmonte, M. K. (2005). "Autism: A Window onto the Development of the Social and the Analytic Brain." *Annu Rev Neurosci* **28**: 109-126.
- Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., and Williams, S. C. (1999). "Social Intelligence in the Normal and Autistic Brain: An Fmri Study." *Eur J Neurosci* **11**(6): 1891-1898.
- Baron, J. (1992). "The Effect of Normative Beliefs on Anticipated Emotions." *Journal of Personality and Social Psychology* **63**: 320-330.
- Baron, J. (2000). *Thinking and Deciding*. Cambridge, Cambridge University Press.
- Berg, J., Dickhaut, J., and McCabe, K. (1995). "Trust, Reciprocity, and Social-History." *Games and Economic Behavior* **10**(1): 122-142.
- Berridge, K. C. (2000). In *The Psychology of Learning and Motivation*. Medin, D. L., Eds. New York, NY, Academic Press: 223-278.
- Bhatt, M., and Camerer, C. F. (2005). "Self-Referential Thinking and Equilibrium as States of Mind in Games: Fmri Evidence." *Games and Economic Behavior* **52**(2): 424-459.
- Blamire, A. M., Ogawa, S., Ugurbil, K., Rothman, D., McCarthy, G., Ellermann, J. M., Hyder, F., Rattner, Z., and Shulman, R. G. (1992). "Dynamic Mapping of the Human Visual Cortex by High-Speed Magnetic Resonance Imaging." *Proc Natl Acad Sci U S A* **89**(22): 11069-11073.
- Borg, J., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). "Consequences, Action, and Intention as Factors in Moral Judgments: An Fmri Investigation." *Journal of Cognitive Neuroscience* **18**(5): 803-817.

- Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., and Cohen, J. D. (1999). "Conflict Monitoring Versus Selection-for-Action in Anterior Cingulate Cortex." *Nature* **402**(6758): 179-181.
- Brandenburger, A. (1996). Strategic and Structural Uncertainty in Games. In *Wise Choices: Games, Decisions, and Negotiations*. Zeckhauser, R., Keeney, R., and Sebenius, J., Eds. Boston, MA, Harvard Business School Press.
- Brune, M. (2005). ""Theory of Mind" In Schizophrenia: A Review of the Literature." *Schizophrenia Bulletin* **31**: 21-42.
- Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., and Decety, J. (2000). "A Pet Investigation of the Attribution of Intentions with a Nonverbal Task." *Neuroimage* **11**(2): 157-166.
- Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., and Rosen, B. R. (1996). "Detection of Cortical Activation During Averaged Single Trials of a Cognitive Task Using Functional Magnetic Resonance Imaging." *Proc Natl Acad Sci U S A* **93**(25): 14878-14883.
- Bush, G., Luu, P., and Posner, M. I. (2000). "Cognitive and Emotional Influences in Anterior Cingulate Cortex." *Trends Cogn Sci* **4**(6): 215-222.
- Cabeza, R., Grady, C. L., Nyberg, L., McIntosh, A. R., Tulving, E., Kapur, S., Jennings, J. M., Houle, S., and Craik, F. I. (1997). "Age-Related Differences in Neural Activity During Memory Encoding and Retrieval: A Positron Emission Tomography Study." *J Neurosci* **17**(1): 391-400.
- Camerer, C. (2003). *Behavioral Game Theory : Experiments in Strategic Interaction*. New York, N.Y., Princeton University Press.
- Camerer, C. F., and Weigelt, K. (1988). "Experimental Tests of a Sequential Equilibrium Reputation Model." *Econometrica* **56**: 1-36.
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., and Cohen, J. D. (1998). "Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance." *Science* **280**(5364): 747-749.
- Carter, C. S., Macdonald, A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., and Cohen, J. D. (2000). "Parsing Executive Processes: Strategic Vs. Evaluative Functions of the Anterior Cingulate Cortex." *Proc Natl Acad Sci U S A* **97**(4): 1944-1948.
- Cavanna, A. E., and Trimble, M. R. (2006). "The Precuneus: A Review of Its Functional Anatomy and Behavioural Correlates." *Brain* **129**(Pt 3): 564-583.
- Cichocki, A., and Amari, S. (2003). *Adaptive Blind Signal and Image Processing*. New York, NY, Wiley & Sons Inc.
- Cohen, J. D., Botvinick, M., and Carter, C. S. (2000). "Anterior Cingulate and Prefrontal Cortex: Who's in Control?" *Nat Neurosci* **3**(5): 421-423.
- Coleman, J. S. (1990). Foundations of Social Theory. Eds. Cambridge, MA, Harvard Univ. Press: 177-179.
- Cover, T., and Thomas, J. (1991). *Elements in Information Theory*. New York, NY, John Wiley & Sons.
- Critchley, H. D., Mathias, C. J., Josephs, O., O'Doherty, J., Zanini, S., Dewar, B. K., Cipolotti, L., Shallice, T., and Dolan, R. J. (2003). "Human Cingulate Cortex and Autonomic Control: Converging Neuroimaging and Clinical Evidence." *Brain* **126**(Pt 10): 2139-2152.
- Dal Bo, P. (2005). "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games." *American Economic Review* **95**: 1591-1604.

- Damasio, A. R. (1994). *Descartes' Error : Emotion, Reason, and the Human Brain*. New York, NY, Putnam.
- Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., and Hichwa, R. D. (2000). "Subcortical and Cortical Brain Activity During the Feeling of Self-Generated Emotions." *Nat Neurosci* **3**(10): 1049-1056.
- David, O., Cosmelli, D., and Friston, K. J. (2004). "Evaluation of Different Measures of Functional Connectivity Using a Neural Mass Model." *Neuroimage* **21**(2): 659-673.
- Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA, The MIT Press.
- de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). "The Neural Basis of Altruistic Punishment." *Science* **305**(5688): 1254-1258.
- Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., and Meltzoff, A. N. (2004). "The Neural Bases of Cooperation and Competition: An Fmri Investigation." *Neuroimage* **23**(2): 744-751.
- Deiniger, K., and Squire, L. (1998). "New Ways of Looking at Old Issues: Inequality and Growth." *Journal of Development Economics* **57**: 259-287.
- Delgado, M. R., Frank, R. H., and Phelps, E. A. (2005). "Perceptions of Moral Character Modulate the Neural Systems of Reward During the Trust Game." *Nat Neurosci* **8**(11): 1611-1618.
- Delgado, M. R., Locke, H. M., Stenger, V. A., and Fiez, J. A. (2003). "Dorsal Striatum Responses to Reward and Punishment: Effects of Valence and Magnitude Manipulations." *Cognitive, Affective, & Behavioral Neuroscience* **3**: 27-38.
- Delgado, M. R., Miller, M. M., Inati, S., and Phelps, E. A. (2005). "An Fmri Study of Reward-Related Probability Learning." *Neuroimage* **24**(3): 862-873.
- Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., and Fiez, J. A. (2000). "Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum." *J Neurophysiol* **84**(6): 3072-3077.
- Deppe, M., Schwindt, W., Kugel, H., Plassmann, H., and Kenning, P. (2005). "Nonlinear Responses within the Medial Prefrontal Cortex Reveal When Specific Implicit Information Influences Economic Decision Making." *J Neuroimaging* **15**(2): 171-182.
- Dickinson, A., and Balleine, B. W. (2002). In *Steven's Handbook of Experimental Psychology*. Gallistel, C. R., Eds. New York, NY, Wiley. **3**: 26-72.
- Dougherty, D. D., Shin, L. M., Alpert, N. M., Pitman, R. K., Orr, S. P., Lasko, M., Macklin, M. L., Fishman, A. J., and Rauch, S. L. (1999). "Anger in Healthy Men: A Pet Study Using Script-Driven Imagery." *Biological Psychiatry* **46**: 466-472.
- Eisenberger, N. I., Lieberman, M. D., and Williams, K. D. (2003). "Does Rejection Hurt? An Fmri Study of Social Exclusion." *Science* **302**(5643): 290-292.
- Elliott, R., Friston, K. J., and Dolan, R. J. (2000). "Dissociable Neural Responses in Human Reward Systems." *J Neurosci* **20**(16): 6159-6165.
- Ellsberg, D. (1961). "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics* **75**(4): 643-669.
- Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., and Peters, T. M. (1993). *3d Statistical Neuroanatomical Models from 305 Mri Volumes*. IEEE-Nuclear Science Symposium and Medical Imaging Conference 1993.

- Fehr, E., and Fischbacher, U. (2003). "The Nature of Human Altruism." *Nature* **425**(6960): 785-791.
- Fehr, E., and Gächter, S. (1998). "Reciprocity and Economics - the Economic Implications of Homo Reciprocans." *European Economic Review* **42**(845-859).
- Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Q. J. Econ.* **108**: 437-459.
- Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). "Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons." *Science* **299**(5614): 1898-1902.
- Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* **7**: 179-188.
- Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., and Frith, C. D. (1995). "Other Minds in the Brain: A Functional Imaging Study Of "Theory of Mind" In Story Comprehension." *Cognition* **57**(2): 109-128.
- Frackowiak, R. S. J., Ashburner, J. T., Penny, W. D., and Zeki, S. (2002). *Human Brain Function*. London, Academic Press.
- Friston, K. J. (2003). Functional Connectivity. In *Human Brain Function*. Frackowiak, R. S. J., Friston, K. J., Frith, C. et al, Eds. San Diego, CA, Academic Press.
- Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998). "Event-Related Fmri: Characterizing Differential Responses." *Neuroimage* **7**(1): 30-40.
- Friston, K. J., Holmes, A. P., Worsley, K., Poline, J. B., Frith, C. D., and Frackowiak, R. S. J. (1995). "Statistical Parametric Maps in Functional Brain Imaging: A General Linear Approach." *Human Brain Mapping* **2**: 189-210.
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., and Turner, R. (1996). "Movement-Related Effects in Fmri Time-Series." *Magnetic Resonance in Medicine* **35**(3): 346-355.
- Frith, C. D., and Frith, U. (1999). "Interacting Minds—a Biological Basis." *Science* **286**(5445): 1692-1695.
- Frith, U., and Frith, C. (2001). "The Biological Basis of Social Interaction." *Current Directions in Psychological Science* **10**: 151-155.
- Fudenberg, D., and Levine, D. (1998). *The Theory of Learning in Games*. Cambridge, MA, MIT Press.
- Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., and Frith, C. D. (2000). "Reading the Mind in Cartoons and Stories: An Fmri Study of 'Theory of Mind' in Verbal and Nonverbal Tasks." *Neuropsychologia* **38**(1): 11-21.
- Gehring, W. J., and Knight, R. T. (2000). "Prefrontal-Cingulate Interactions in Action Monitoring." *Nat Neurosci* **3**(5): 516-520.
- Georgieff, N., and Jeannerod, M. (1998). "Beyond Consciousness of External Reality: A "Who" System for Consciousness of Action and Self-Consciousness." *Conscious Cogn* **7**(3): 465-477.
- Gini, C. (1912). Variabilità E Mutabilità. In *Reprinted in Memorie Di Metodologica Statistica* (1955). E., P., and Salvemini, T., Eds. Rome, Libreria Eredi Virgilio Veschi.
- Gini, C. (1921). "Measurement of Inequality and Incomes." *The Economic Journal* **31**: 124-126.
- Glimcher, P. W. (2003). *Decisions, Uncertainty, and the Brain : The Science of Neuroeconomics*. Cambridge, Mass., MIT Press.
- Glimcher, P. W., and Rustichini, A. (2004). "Neuroeconomics: The Consilience of Brain and Decision." *Science* **306**(5695): 447-452.

- Goebel, B., Dawy, Z., Hagenauer, J., and Mueller, J. C. (2005). An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates. *IEEE International Conference on Communications (ICC 2005)*. Seoul, South Korea: 1102-1106.
- Goel, V., Grafman, J., Sadato, N., and Hallett, M. (1995). "Modeling Other Minds." *Neuroreport* **6**(13): 1741-1746.
- Goldman-Rakic, P. S. (1996). "The Prefrontal Landscape: Implications of Functional Architecture for Understanding Human Mentation and the Central Executive." *Philos Trans R Soc Lond B Biol Sci* **351**(1346): 1445-1453.
- Greene, J. (2003). "From Neural 'Is' to Moral 'Ought': What Are the Moral Implications of Neuroscientific Moral Psychology?" *Nat Rev Neurosci* **4**(10): 846-849.
- Greene, J., and Haidt, J. (2002). "How (and Where) Does Moral Judgment Work?" *Trends Cogn Sci* **6**(12): 517-523.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* **44**(2): 389-400.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). "An Fmri Investigation of Emotional Engagement in Moral Judgment." *Science* **293**(5537): 2105-2108.
- Guimaraes, A. R., Melcher, J. R., Talavage, T. M., Baker, J. R., Ledden, P., Rosen, B. R., Kiang, N. Y., Fullerton, B. C., and Weisskoff, R. M. (1998). "Imaging Subcortical Auditory Activity in Humans." *Hum Brain Mapp* **6**(1): 33-41.
- Gul, F., and Pesendorfer, W. (2005). "The Case for Mindless Economics." *Princeton University Working Paper*.
- Hamilton, W. D. (1964). "The Genetical Evolution of Social Behaviour. I." *J Theor Biol* **7**(1): 1-16.
- Hamilton, W. D. (1964). "The Genetical Evolution of Social Behaviour. II." *J Theor Biol* **7**(1): 17-52.
- Haruno, M., and Kawato, M. (2006). "Different Neural Correlates of Reward Expectation and Reward Expectation Error in the Putamen and Caudate Nucleus During Stimulus-Action-Reward Association Learning." *J Neurophysiol* **95**(2): 948-959.
- Heeger, D. J., Huk, A. C., Geisler, W. S., and Albrecht, D. G. (2000). "Spikes Versus Bold: What Does Neuroimaging Tell Us About Neuronal Activity?" *Nat Neurosci* **3**(7): 631-633.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., and Tracer, D. (2005). "'Economic Man' In Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behav Brain Sci* **28**(6): 795-815; discussion 815-755.
- Hill, E. L., and Frith, U. (2003). "Understanding Autism: Insights from Mind and Brain." *Philos Trans R Soc Lond B Biol Sci* **358**(1430): 281-289.
- Holroyd, C. B., and Coles, M. G. (2002). "The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity." *Psychol Rev* **109**(4): 679-709.
- Holroyd, C. B., Nieuwenhuis, S., Yeung, N., and Cohen, J. D. (2003). "Errors in Reward Prediction Are Reflected in the Event-Related Brain Potential." *Neuroreport* **14**(18): 2481-2484.

- Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). "Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making." *Science* **310**(5754): 1680-1683.
- Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, Mass., Sinauer Associates, Publishers.
- Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., and Platt, M. L. (2006). "Neural Signatures of Economic Preferences for Risk and Ambiguity." *Neuron* **49**(5): 765-775.
- Hyvarinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. New York, NY, Wiley & Sons Inc.
- Jezzard, P. (2001). *Functional Mri: An Introduction to Methods*. New York, NY, Oxford University Press.
- Johns, L. C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., and McGuire, P. K. (2001). "Verbal Self-Monitoring and Auditory Verbal Hallucinations in Patients with Schizophrenia." *Psychol Med* **31**(4): 705-715.
- Johnson, P. A., Hurley, R. A., Benkelfat, C., Herpertz, S. C., and Taber, K. H. (2003). "Understanding Emotion Regulation in Borderline Personality Disorder: Contributions of Neuroimaging." *J Neuropsychiatry Clin Neurosci* **15**(4): 397-402.
- Kamitani, Y., and Tong, F. (2005). "Decoding the Visual and Subjective Contents of the Human Brain." *Nat Neurosci* **8**(5): 679-685.
- Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., and Heatherton, T. F. (2002). "Finding the Self? An Event-Related Fmri Study." *J Cogn Neurosci* **14**(5): 785-794.
- Keynes, J. M. (1921). *A Treatise on Probability*. Eds. London, Macmillan: pp. 75-76, 315.
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange." *Science* **308**(5718): 78-83.
- Knutson, B., and Cooper, J. C. (2005). "Functional Magnetic Resonance Imaging of Reward Prediction." *Curr Opin Neurol* **18**(4): 411-417.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio, A. (2007). "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements." *Nature* **446**(7138): 908-911.
- Konow, J. (2003). "Which Is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature* **XLI**: 1188-1239.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). "Oxytocin Increases Trust in Humans." *Nature* **435**(7042): 673-676.
- Kuhnen, C. M., and Knutson, B. (2005). "The Neural Basis of Financial Risk Taking." *Neuron* **47**(5): 763-770.
- Lee, K. H., Farrow, T. F., Spence, S. A., and Woodruff, P. W. (2004). "Social Cognition, Brain Networks and Schizophrenia." *Psychol Med* **34**(3): 391-400.
- Lieberman, M. D., Jarcho, J. M., and Satpute, A. B. (2004). "Evidence-Based and Intuition-Based Self-Knowledge: An Fmri Study." *J Pers Soc Psychol* **87**(4): 421-435.
- Lieberman, M. D., and Pfeifer, J. H. (2005). In *Cognitive Neuroscience of Emotional and Social Behavior*. Easton, A., and Emery, N., Eds. Philadelphia, Psychology Press: 195-235.
- Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). "Neurophysiological Investigation of the Basis of the Fmri Signal." *Nature* **412**(6843): 150-157.
- Maynard Smith, J., and Price, G. R. (1973). "The Logic of Animal Conflict." *Nature* **146**: 15-18.

- McCabe, K. (2003). Neuroeconomics. In *Encyclopedia of Cognitive Sciences*. Nature Publishing Group, M. P., Eds. New York, NY, Nadel, L.: 294-298.
- McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange." *Proceedings of the National Academy of Sciences of the United States of America* **98**(20): 11832-11835.
- McClure, E. B., Parrish, J. M., Nelson, E. E., Easter, J., Thorne, J. F., Rilling, J. K., Ernst, M., and Pine, D. S. (2007). "Responses to Conflict and Cooperation in Adolescents with Anxiety and Mood Disorders." *J Abnorm Child Psychol*.
- McClure, S. M., Berns, G. S., and Montague, P. R. (2003). "Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum." *Neuron* **38**(2): 339-346.
- Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K. (1999). *Fisher Discriminant Analysis with Kernels*. IEEE Neural Networks for Signal Processing Workshop 1999.
- Milham, M. P., Banich, M. T., Claus, E. D., and Cohen, N. J. (2003). "Practice-Related Effects Demonstrate Complementary Roles of Anterior Cingulate and Prefrontal Cortices in Attentional Control." *Neuroimage* **18**(2): 483-493.
- Moddemeijer, R. (1989). "On Estimation of Entropy and Mutual Information of Continuous Distributions." *Signal Processing* **16**: 233-248.
- Moll, J., de Oliveira-Souza, R., Bramati, I. E., and Grafman, J. (2002). "Functional Networks in Emotional Moral and Nonmoral Social Judgments." *Neuroimage* **16**: 696-703.
- Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., and Pessoa, L. (2002). "The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions." *J Neurosci* **22**(7): 2730-2736.
- Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). "Opinion: The Neural Basis of Human Moral Cognition." *Nat Rev Neurosci* **6**(10): 799-809.
- Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., Wiest, M. C., Karpov, I., King, R. D., Apple, N., and Fisher, R. E. (2002). "Hyperscanning: Simultaneous Fmri During Linked Social Interactions." *Neuroimage* **16**(4): 1159-1164.
- Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. J. (1995). "Bee Foraging in Uncertain Environments Using Predictive Hebbian Learning." *Nature* **377**(6551): 725-728.
- Montague, P. R., Hyman, S. E., and Cohen, J. D. (2004). "Computational Roles for Dopamine in Behavioural Control." *Nature* **431**(7010): 760-767.
- Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). "Midbrain Dopamine Neurons Encode Decisions for Future Action." *Nat Neurosci* **9**(8): 1057-1063.
- Nelder, J. A., and Mead, R. (1965). "A Simplex-Method for Function Minimization." *Computer Journal* **7**(4): 308-313.
- Nielsen, F. A., Balslev, D., and Hansen, L. K. (2005). "Mining the Posterior Cingulate: Segregation between Memory and Pain Components." *Neuroimage* **27**(3): 520-532.
- Nowak, M. A., and Sigmund, K. (1992). "Tit-for-Tat in Heterogeneous Populations." *Nature* **355**: 250-253.
- O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning." *Science* **304**(5669): 452-454.

- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). "Temporal Difference Models and Reward-Related Learning in the Human Brain." *Neuron* **38**(2): 329-337.
- O'Reilly, R. C., Braver, T. S., and Cohen, J. D. (1999). In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Miyake, A., and Shah, P., Eds. New York, NY, Cambridge Univ. Press: 375-411.
- Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., and Mackey, S. C. (2004). "Reflecting Upon Feelings: An Fmri Study of Neural Systems Supporting the Attribution of Emotion to Self and Other." *J Cogn Neurosci* **16**(10): 1746-1772.
- Ochsner, K. N., Kosslyn, S. M., Cosgrove, G. R., Cassem, E. H., Price, B. H., Nierenberg, A. A., and Rauch, S. L. (2001). "Deficits in Visual Cognition and Attention Following Bilateral Anterior Cingulotomy." *Neuropsychologia* **39**(3): 219-230.
- Ogawa, S., Lee, T. M., Nayak, A. S., and Glynn, P. (1990). "Oxygenation-Sensitive Contrast in Magnetic-Resonance Image of Rodent Brain at High Magnetic-Fields." *Magnetic Resonance in Medicine* **14**(1): 68-78.
- Ollinger, J. M., Shulman, G. L., and Corbetta, M. (2001). "Separating Processes within a Trial in Event-Related Functional Mri." *Neuroimage* **13**(1): 210-217.
- Otten, L. J., and Rugg, M. D. (2001). "Task-Dependency of the Neural Correlates of Episodic Encoding as Measured by Fmri." *Cereb Cortex* **11**(12): 1150-1160.
- Pagnoni, G., Zink, C. F., Montague, P. R., and Berns, G. S. (2002). "Activity in Human Ventral Striatum Locked to Errors of Reward Prediction." *Nat Neurosci* **5**(2): 97-98.
- Paulus, M. P., Hozack, N., Frank, L., and Brown, G. G. (2002). "Error Rate and Outcome Predictability Affect Neural Activation in Prefrontal Cortex and Anterior Cingulate During Decision-Making." *Neuroimage* **15**(4): 836-846.
- Petrides, M. (1994). Frontal Lobes and Working Memory: Evidence from Investigation of the Effects of Cortical Excisions in Nonhumans Primates. In *Handbook of Neuropsychology*. Boller, F., and Grafman, J., Eds. Amsterdam, Elsevier.
- Phan, K. L., Liberzon, I., Welsh, R. C., Britton, J. C., and Taylor, S. F. (2003). "Habituation of Rostral Anterior Cingulate Cortex to Repeated Emotionally Salient Pictures." *Neuropsychopharmacology* **28**(7): 1344-1350.
- Premack, D., and Woodruff, G. (1978). "Does the Chimpanzee Have a Theory of Mind?" *The Behavioral and Brain Sciences* **4**: 515-526.
- Preuschoff, K., Bossaerts, P., and Quartz, S. R. (2006). "Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures." *Neuron* **51**(3): 381-390.
- Quenouille, M. H. (1956). "Notes on Bias in Estimation." *Biometrika* **43**: 353-360.
- Rachlin, H. (2002). "Altruism and Selfishness." *Behav Brain Sci* **25**(2): 239-250; discussion 251-296.
- Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., and Bushnell, M. C. (1997). "Pain Affect Encoded in Human Anterior Cingulate but Not Somatosensory Cortex." *Science* **277**: 968-971.
- Remijne, P. L., Nielen, M. M., Uylings, H. B., and Veltman, D. J. (2005). "Neural Correlates of a Reversal Learning Task with an Affectively Neutral Baseline: An Event-Related Fmri Study." *Neuroimage* **26**(2): 609-618.
- Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). "A Neural Basis for Social Cooperation." *Neuron* **35**(2): 395-405.

- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2004). "The Neural Correlates of Theory of Mind within Interpersonal Interactions." *Neuroimage* **22**(4): 1694-1703.
- Ritov, I., and Baron, J. (1992). "Status-Quo and Omission Bias." *Journal of Risk and Uncertainty* **5**: 49-61.
- Rolls, E. T. (1996). "The Orbitofrontal Cortex." *Philos Trans R Soc Lond B Biol Sci* **351**(1346): 1433-1443; discussion 1443-1434.
- Roulston, M. S. (1999). "Estimating the Errors on Measured Entropy and Mutual Information." *Physica D* **125**: 285-294.
- Roy, C. S., and Sherrington, C. S. (1890). "On the Regulation of the Blood Supply of the Brain." *Journal of Physiology* **11**: 85-108.
- Sandrini, M., Cappa, S. F., Rossi, S., Rossini, P. M., and Miniussi, C. (2003). "The Role of Prefrontal Cortex in Verbal Episodic Memory: Rtms Evidence." *J Cogn Neurosci* **15**(6): 855-861.
- Sanfey, A. G., Loewenstein, G., McClure, S. M., and Cohen, J. D. (2006). "Neuroeconomics: Cross-Currents in Research on Decision-Making." *Trends in Cognitive Sciences* **10**(3): 108-116.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science* **300**(5626): 1755-1758.
- Schokkaert, E., and Overlaet, B. (1989). "Moral Intuitions and Economic Models of Distributive Justice." *Social Choice and Welfare* **6**: 19-31.
- Schultz, W., Dayan, P., and Montague, P. R. (1997). "A Neural Substrate of Prediction and Reward." *Science* **275**(5306): 1593-1599.
- Seber, G. A. (2004). *Multivariate Observations*. New York, NY, Wiley.
- Seger, C. A., Stone, M., and Keenan, J. P. (2004). "Cortical Activations During Judgments About the Self and an Other Person." *Neuropsychologia* **42**(9): 1168-1177.
- Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). "Temporal Difference Models Describe Higher-Order Learning in Humans." *Nature* **429**(6992): 664-667.
- Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). "Opponent Appetitive-Aversive Neural Processes Underlie Predictive Learning of Pain Relief." *Nat Neurosci* **8**(9): 1234-1240.
- Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*(27): 379-423 & 623-656.
- Shmuel, A., Augath, M., Oeltermann, A., and Logothetis, N. K. (2006). "Negative Functional Mri Response Correlates with Decreases in Neuronal Activity in Monkey Visual Area V1." *Nat Neurosci* **9**(4): 569-577.
- Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). "Empathic Neural Responses Are Modulated by the Perceived Fairness of Others." *Nature* **439**(7075): 466-469.
- Spranca, M., Minsk, E., and Baron, J. (1991). "Omission and Commission in Judgment and Choice." *Journal of Experimental Social Psychology* **27**: 76-105.
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA, MIT Press.

- Tanaka, S., Honda, M., and Sadato, N. (2005). "Modality-Specific Cognitive Function of Medial and Lateral Human Brodmann Area 6." *J Neurosci* **25**(2): 496-501.
- Trivers, R. L. (1971). "The Evolution of Reciprocal Altruism." *Q. Rev. Biol.* **46**: 35-57.
- Tsai, A., J.W., F., Wible, C., Wells, W. M., Kim, J., and Willsky, A. S. (1999). Analysis of Functional Mri Data Using Mutual Information. *Medical Image Computing and Computer-Assisted Intervention*. Cambridge, UK.
- van Leijenhorst, L., Crone, E. A., and Bunge, S. A. (2006). "Neural Correlates of Developmental Differences in Risk Estimation and Feedback Processing." *Neuropsychologia* **44**(11): 2158-2170.
- Varian, H. R. (1975). "Distributive Justice, Welfare Economics, and the Theory of Fairness." *Philosophy and Public Affairs* **4**(3): 223-247.
- Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., and Zilles, K. (2001). "Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective." *Neuroimage* **14**(1 Pt 1): 170-181.
- Vogeley, K., and Fink, G. R. (2003). "Neural Correlates of the First-Person-Perspective." *Trends Cogn Sci* **7**(1): 38-42.
- Volz, K. G., Schubotz, R. I., and von Cramon, D. Y. (2003). "Predicting Events of Varying Probability: Uncertainty Investigated by Fmri." *Neuroimage* **19**(2 Pt 1): 271-280.
- Weissman, D. H., Giesbrecht, B., Song, A. W., Mangun, G. R., and Woldorff, M. G. (2003). "Conflict Monitoring in the Human Anterior Cingulate Cortex During Selective Attention to Global and Local Object Features." *Neuroimage* **19**(4): 1361-1368.
- Wicker, B., Perrett, D. I., Baron-Cohen, S., and Decety, J. (2003). "Being the Target of Another's Emotion: A Pet Study." *Neuropsychologia* **41**: 139-146.
- Wittman, D. (1984). "The Geometry of Justice: Three Existence and Uniqueness Theorems." *Theory and Decision* **16**: 239-250.
- Yaari, M. E., and Bar-Hillel, M. (1984). "On Dividing Justly." *Social Choice and Welfare* **1**: 1-24.